Al-Quds University
Deanship of Graduate Studies
Faculty of Health Profession
Medical Imaging Technology

A Hybrid Artificial Intelligence Approach for Early Detection of Breast Cancer and Classification from Mammogram Images in Palestine

Omar Faiq Sadeq Daraghmeh

M.Sc. Thesis

Jerusalem - Palestine
1445/ 2024

A Hybrid Artificial Intelligence Approach for Early Detection of Breast Cancer and Classification from Mammogram Images in Palestine

Prepared By:

Omar Faiq Sadeq Daraghmeh

Supervisor: **Dr. Radwan Qasrawi**

This Thesis submitted in partial fulfillment of requirements for the degree of Master of Medical Imaging Technology Faculty of Graduate studies - Al-Quds University

1445/2024

Al-Quds University

Deanship of Graduate Studies

Faculty of Health Profession

Medical Imaging Technology

Thesis Approval

A Hybrid Artificial Intelligence Approach for Early Detection of Breast Cancer and Classification from Mammogram Images in Palestine

Prepared by: **Omar Faiq Sadeq Daraghmeh**

Registration No: **22010988**

Supervisor: **Dr. Radwan Qasrawi**

The master's thesis was submitted and accepted on 26/05/2024.

The names and signatures of the examining committee members are as follows:

Head of Committee: **Dr. Radwan Qasrawi**	Signature:
Internal Examiner: **Dr. Hussein Al-Masri**	Signature:
External Examiner: **Dr. Derar Eleyan**	Signature:

Jerusalem – Palestine

1445 / 2024

Dedication

This work is humbly dedicated to the resilient and steadfast people of Gaza, whose unwavering determination in the face of adversity and aggression from the Israeli occupiers is an inspiration to all humanity. Despite enduring years of military bombardments, crippling blockades, and unconscionable violations of human rights, the proud people of Gaza remain defiant, their resilience unbroken, their hopes for freedom, justice and self-determination undiminished. As researchers, we stand in solidarity with the Palestinian struggle against oppression and colonial occupation. This study is a small contribution towards the development of improved diagnostic tools that we hope can one day benefit all people, including those in Gaza whose access to healthcare has been so cruelly impeded. The pursuit of knowledge and scientific inquiry must never be dehumanized or cynically decoupled from the harsh realities faced by oppressed peoples around the world. We dedicate our efforts to the people of Gaza who have demonstrated, through their tenacity and resilience, the boundless capacities of the human spirit. In the words of the revered Palestinian poet Mahmoud Darwish: "We have on this earth what makes life worth living." Our research endeavors are inspired by your courage and your just struggle - may the fruits of our labor aid in the creation of a more dignified, equitable and peaceful future for all. To the people of Gaza, we salute your heroic resistance. Your cause is the cause of all those who believe in freedom, human rights and the fundamental dignity of all people. You will always remain the conscience that guides our scientific pursuits.

Declaration:

I certify that this thesis submitted for the degree of Master is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or Institution.

Omar Faiq Sadeq Daraghmeh

Signed: 

Date: 26/05/2024

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Dr. Radwan Qasrawi, my research supervisor, for his exceptional guidance, unwavering support, and invaluable mentorship throughout this research endeavor. Dr. Qasrawi's profound knowledge, insightful feedback, and constant encouragement have been instrumental in successfully completing this study, and I am truly grateful for the countless hours he dedicated to discussing ideas, reviewing my work, and providing constructive criticism. His guidance pushed me to strive for excellence. I am also deeply indebted to the esteemed members of my thesis committee, Dr. Hussein Al-masri and Dr. Derar Eleyan, whose valuable suggestions, thought-provoking questions, and constructive critiques significantly enriched the quality of this research and broadened my understanding of the subject matter. Their expertise and diverse perspectives played a crucial role in shaping this study.

Furthermore, I would like to extend my sincere appreciation to the Department of Medical Imaging Technology at Al-Quds University for providing state-of-the-art facilities, computational resources, and a stimulating research environment that enabled me to seamlessly conduct this study. My colleagues and fellow researchers at the Al-Quds Business Center for Innovation, Technology & Entrepreneurship, especially Sulieman Thwib, deserve my gratitude for their collaboration, stimulating discussions, and invaluable feedback. Their contributions significantly refined the ideas and methodologies presented in this work, fostering an environment of intellectual growth and camaraderie. I would also like to acknowledge the Dunya Women's Cancer Center, particularly Dr. Nufuz Maslamani, Dr. Haneen Owienah and Dr. Muath Melhem, for providing access to the invaluable mammogram image dataset that formed the foundation of this research. Their commitment to advancing scientific knowledge and facilitating collaborations has been instrumental in driving progress in the field of breast cancer diagnosis.

Finally, I am profoundly grateful to my family and friends for their unconditional love, support, and unwavering belief in me. To my parents, Faiq and Ferial Daraghmeh, thank you for instilling in me the values of perseverance, resilience, and the pursuit of knowledge. Your constant encouragement and sacrifices have laid the foundation of my success. To my partner, Ibraheem Qdaih, your patience, understanding, and unwavering support during the most challenging times have been a constant source of strength and motivation. And to my friends, Sohaib Daraghmeh and Fadi Abu Amer, thank you for being my support system, offering a listening ear, and providing

much-needed laughter and respite throughout this journey. To all those who have contributed directly or indirectly to the successful completion of this research, I express my heartfelt gratitude. Your support, guidance, and contributions have been invaluable, and I am forever indebted to you for helping me utilize your materials and facilities for the progress and achievements of this work.

Abstract

Breast cancer is a significant global health concern, especially in Palestine, and early and accurate diagnosis is crucial for improving patient outcomes and survival rates. However, despite advancements in medical technology and screening techniques, missed diagnoses remain a persistent challenge in breast cancer detection. This study investigates the use of hybrid artificial intelligence (AI) models that combine deep learning and machine learning techniques to predict benign and malignant breast cancer from mammogram images. The study starts by utilizing pre-trained convolutional neural network models, namely VGG16 and DenseNet121, for feature extraction from mammogram images. These deep learning models have been trained on large datasets and have learned to identify various patterns and features within images. By extracting these features from mammograms, the models can capture important information that is relevant to the classification of breast cancer. The extracted features are then used to train several machine learning classifiers, including logistic regression, support vector machines, random forests, and gradient boosting models. These classifiers learn to recognize patterns and make predictions based on the extracted features.

To evaluate the performance of the hybrid AI models, the study is conducted in three stages. In the first stage, the original mammogram images are used for classification. In the second stage, the mammogram images are enhanced using various image preprocessing and enhancement techniques. Finally, in the third stage, the models are tested on new mammogram images to assess their generalization capabilities. To enhance the mammogram images, several image processing techniques are applied. These include morphological erosion preprocessing, Contrast-Limited Adaptive Histogram Equalization (CLAHE), Laplacian of Gaussian (LoG) edge enhancement, and unsharp masking. These techniques aim to improve the visibility of important structures and features within the images, making it easier for the AI models to make accurate predictions.

In the second stage, when predicting benign cases from the enhanced mammogram images, the logistic regression classifier with DenseNet121 features achieves remarkable performance. It achieves the highest accuracy of 0.991, precision of 0.996, F1-score of 0.989, and an AUC of 0.999. The support vector machine with DenseNet121 features also performs well, with an accuracy of 0.986 and an AUC of 0.999. The logistic regression model with VGG16 features demonstrates the fastest predictive time, requiring only 0.13 seconds. Similarly, in predicting

malignant cases from the enhanced images, the logistic regression classifier with DenseNet121 features excels with the highest accuracy of 0.995, precision of 0.995, recall of 0.995, F1-score of 0.995, and an AUC of 0.999. The support vector machine with DenseNet121 features follows closely with an accuracy of 0.992 and an AUC of 0.998. The logistic regression model with VGG16 features maintains its fast predictive time, taking only 0.08 seconds. The study demonstrates that the enhanced mammogram images in the second stage consistently outperform the original and new test images in the first and third stages, respectively. This emphasizes the significant impact of image preprocessing and enhancement techniques on the predictive capabilities of the hybrid AI models. The findings highlight the potential of combining deep learning for feature extraction and machine learning for classification in achieving high accuracy, precision, recall, F1-scores, and AUC values for predicting breast cancer malignancy from mammogram images.

In conclusion, the study demonstrates the potential of hybrid AI models that combine deep learning and machine learning techniques for the prediction of benign and malignant breast cancer from mammogram images. The integration of deep learning for feature extraction and machine learning for classification, along with image preprocessing and enhancement, results in improved accuracy and performance. These advancements have the potential to enhance breast cancer detection, ultimately leading to better patient outcomes and survival rates.

Table of Contents

List of Figures ..... **X**

List of Tables ..... **XIII**

List of Equations ..... **XIV**

Abbreviations ..... **XV**

Chapter One : Introduction **1**

1.1 Background 1

1.1.1 Breast Cancer Statistics in Palestine 1

1.1.2 Detection and Classification of Breast Cancer 6

1.1.3 The Implementation of Artificial Intelligence (AI) in Breast Imaging 12

1.2 Problem Statement 14

1.3 Research Objectives 16

1.4 Research Questions 16

1.5 Research Justifications 17

1.6 Research Hypotheses 18

Chapter Two: Literature Review **20**

2.1 Machine Learning Applications in Breast Cancer Detection and Classification 20

2.2 Deep Learning Applications in Breast Cancer Detection and Classification 28

2.3 Applications of Hybrid Artificial Intelligence Learning in Breast Cancer Detection and Classification 42

Chapter Three: Methodology **45**

3.1 Ethical statement and confidentiality 46

3.2 Data Collection 46

3.3 Mammography Equipment and Tru-Cut Biopsy Procedure 48

3.4 Data Sorting 51

3.5 Inclusion and Exclusion Criteria for Collected Mammograms 52
3.6 Data Manipulation 53
3.7 Data Preprocessing..... 55
3.8 Dataset..... 59
3.9 Building a Hybrid Artificial Intelligence (AI) Model..... 61
3.9.1 Deep Learning-Based Feature Extraction..... 61
3.9.2 Machine Learning-Based Feature Classification..... 64
3.10 Approaches for Evaluating a Hybrid Artificial Intelligence (AI) Model 68
3.10.1 The Confusion Matrix 69
3.10.2 The Accuracy..... 70
3.10.3 The Receiver Operating Characteristic (ROC)..... 71
3.10.4 The Area under the ROC Curve (AUC) 72
3.10.5 The Precision 73
3.10.6 The F1-Score 73
3.10.7 The Recall..... 74
3.10.8 The Time..... 74
Chapter Four: Results 76
3.1 Contrast-Limited Adaptive Histogram Equalization (CLAHE) 76
3.1 Performance of Hybrid Artificial Intelligence (AI) Models without Enhancement techniques 78
3.2 Performance of Hybrid Artificial Intelligence (AI) Models with Enhancement techniques .. 83
3.3 Validation of Hybrid Artificial Intelligence (AI) Models with Enhancement techniques..... 88
Chapter Five: Discussion..... 93
4.1 Analysis of Contrast-Limited Adaptive Histogram Equalization (CLAHE)..... 93
4.2 Evaluation of hybrid artificial intelligence (AI) models..... 94

4.2.1 Training Stage 1: From Original Mammogram Images 95
4.2.2 Training Stage 2: From Enhanced Mammogram Images 107
4.2.3 Application Stage: From New Mammogram Images 115
4.2.4 Comparison with previous studies 122
Chapter Six: Conclusion 126
6.1 Strength and Limitations 128
6.2 Recommendations 128
6.3 Future study 129
References 130

List of Figures

Figure 1: Progression of Breast Cancer (Saint John's Cancer Institute, 2024). 2
Figure 2: Distribution of Percentage of Top Ten Reported Cancers in all population, Palestine 2022 (PHIC, 2023). 2
Figure 3: Incidence Rate of Top Ten Reported Cancers per 100,000 of population, West Bank, Palestine 2022 (PHIC, 2023). 3
Figure 4: Incidence rates per 100,000 population for the top ten reported cancer types among males and females in the West Bank, Palestine, in 2022 (PHIC, 2023). 3
Figure 5: Proportional Distribution of the most Reported Cancer Deaths of all Reported Cancer Deaths, West Bank, 2022 (PHIC, 2023). 4
Figure 6: A global map representing the incidence and mortality rates of breast cancer, World Health Organization in 2022 (WHO, 2022a). 5
Figure 7: Breast mammograms (MD) taken from two different positions (Justaniah et al., 2022). 6
Figure 8: Imaging Methods for Breast Cancer Detection: Ultrasound (A), Magnetic Resonance Imaging (MRI) (B), Computed Tomography, and PET Scan (C). 7
Figure 9: Graphs depicting three patterns of Kinetic curves typically seen in breast lesions, as intensity enhancement as a function of time. A) Type I – (persistent enhancing), B) Type II – (plateau) and C) Type III – (washout) (Craciunescu et al., 2009). 11
Figure 10: The methodological approach underpinning the research design employed in this study. 46
Figure 11: An image showcasing the Mammomat Revelation device (Healthineers, 2023). 50
Figure 12: The procedure for obtaining a sample using the tru-cut technique (TEAM, 2018). .. 51
Figure 13: The outcomes obtained from employing augmentation methods on mammogram images, encompassing rotations and flips techniques..... 54
Figure 14: The effectiveness of CLAHE algorithm in enhancing benign and malignant mammography images at different clip limits (0, 2, 3, and 4). 57
Figure 15: The process of dividing data into separate subsets for cross-validation, which are used for training and evaluating artificial intelligence models (Duran-Lopez et al., 2020). 60
Figure 16: VGG16 model architecture. 63

Figure 17: DenseNet121 model architecture. 64
Figure 18: Random Forest (RF) (Rybiałek & Jeleń, 2020). 66
Figure 19: Gradient Boosting (GB) (Rybiałek & Jeleń, 2020). 66
Figure 20: Support Vector Machine (SVM) (Azar & El-Said, 2014). 67
Figure 21: Logistic Regression (LR) (Ayer et al., 2010). 67
Figure 22: The Python environment leveraged to construct the code-base for the hybrid models investigated in this research, along with the various software libraries utilized. 68
Figure 23: A Confusion Matrix, a tabular representation that illustrates the correlation between the actual values and the predicted values, highlighting the correct predictions as well as the errors in the predictive model. (Mokhtari et al., 2021) 70
Figure 24: The Receiver Operating Characteristic (ROC) curve, a graphical plot that depicts the trade-off between the true positive rate and the false positive rate, providing a comprehensive evaluation of the forecasting model's performance (Soltani et al., 2019)..... 72
Figure 25: The interplay between Peak Signal-to-Noise Ratio (PSNR) and the Exponential Mean Error (EME) metrics, in conjunction with the ClipLimit parameter, to determine the optimal value for applying Contrast Limited Adaptive Histogram Equalization (CLAHE) in image enhancement. 77
Figure 26: A comparative evaluation of the VGG16 and DenseNet121 feature extraction techniques, in conjunction with all the proposed classification models, for the task of predicting benign breast cancer cases. 80
Figure 27: A comparative evaluation of the predictive performance exhibited by all the proposed classification models when employing the vgg16 feature extractor versus the densenet121 feature extractor for the task of identifying malignant breast cancer cases. 82
Figure 28: A comparative evaluation of the predictive performance achieved by all the proposed classification models when identifying benign breast cancer cases, contrasting the results 85
Figure 29: A comparative evaluation of the predictive performance achieved by all classification models when identifying malignant breast cancer cases, contrasting the results obtained using the VGG16 feature extractor against those obtained with the DenseNet121 feature extractor, after applying image enhancement techniques to the input data. 87
Figure 30: A comparative analysis of the predictive performance achieved by all classification models when identifying benign breast cancer cases from new, unseen mammogram images,

contrasting the results obtained using the VGG16 feature extractor against those obtained with the DenseNet121 feature extractor. 90
Figure 31: A comparative analysis of the predictive performance achieved by all classification models when identifying malignant breast cancer cases from new, unseen mammogram images, contrasting the results obtained using the VGG16 feature extractor against those obtained with the DenseNet121 feature extractor. 92

List of Tables

Table 1: Distribution of Reported Cancer Deaths by Site & Sex, West Bank 2022 (PHIC, 2023). 4
Table 2: Comparison of different types of breast imaging (Zhu et al., 2023)..... 8
Table 3: Breast Imaging Reporting and Data System (BI-RADS) categories (Smithuis, 2014). 10
Table 4: A compilation of prior research studies discussing the utilization of artificial intelligence for the detection and classification of breast cancer from mammogram images, along with the performance metrics attained in each study..... 37
Table 5: The distribution of the number of cases and biopsies gathered between 2018 and 2023 for constructing the database. 47
Table 6: Mammomat Revelation device components and features (Healthineers, 2023). 48
Table 7: The distribution of the number of cases and images following classification by a radiologist, based on the biopsy reports associated with each case..... 52
Table 8: Types of data and augmentation parameter. 54
Table 9: EME and PSNR values for data corresponding to different contrast thresholds..... 77
Table 10: Other state-of-the-art hybrid models reported in previous. 122
Table 11: Performance Evaluation of the Original Data (Before image enhancement) Using Different Machine Learning Classifiers, with Respect to Two Deep Learning Models (VGG16, DenseNet121). 143
Table 12: Performance Evaluation of the Enhanced Data Using CLAHE Algorithm, Using Different Machine Learning Classifiers, with Respect to Two Deep Learning Models i.e. (VGG16, DenseNet121). 144
Table 13: Performance Evaluation of the Proposed Hybrid Model on a Real-World Clinical Dataset with Enhancement (400 benign and 400 Malignant) cases, Referring to Confirmed Histopathology..... 145

List of Equations

$CDF(x) = \sum_{y=0}^{x} \frac{H(y)}{NT}$ 56

$EME = \frac{1}{K_1 K_2} \sum_{L=1}^{K_2} \sum_{K=1}^{K_1} 20\, log \left(\frac{I_{max}(k,l)}{I_{min}(k,l)} \right)$ 57

$PSNR(f,g) = 10\, log_{10} \left(\frac{255^2}{MSE(f,g)} \right)$ 58

$MSE(f,g) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(f_{ij} - g_{ij} \right)^2$ 58

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ 70

$True\ Positive\ Rate\ (TPR)(Sensitivity)(Recall) = \frac{TP}{FN+TP} = \frac{TP}{P}$ 71

$False\ Positive\ Rate\ (FPR)(False\ Alarm\ Rate) = \frac{FP}{TN+FP} = \frac{FP}{N}$ 71

$True\ Negative\ Rate(TNR)(Specificity) = \frac{TN}{TN+FP} = \frac{TN}{N} = 1 - FPR$ 71

$False\ Negative\ Rate\ (FNR) = \frac{FN}{FN+TP} = \frac{FN}{P}$ 71

$Precision = \frac{TP}{TP+FP}$ 73

$F1\ Score = \frac{2 \times Precisiqon \times Recall}{Precision+Recall}$ 74

Abbreviations

ACR:	American College of Radiology
ACS:	American Cancer Society
AEC:	Automatic Exposure Control
AHE:	Adaptive Histogram Equalization
AI:	Artificial Intelligence
ANN:	Artificial Neural Networks
AUC:	Area Under the ROC Curve
BI-RADS:	Breast Imaging Reporting & Data System
BP:	BackPropagation
CAD:	Computer-Aided Detection
CC:	CranioCaudal
CDF:	Cumulative Distribution Function.
CEDM:	Contrast-Enhanced Digital Mammography
CFS	Correlation-based Feature Selection
CLAHE:	Contrast-Limited Adaptive Histogram Equalization
CNN:	Convolutional Neural Networks
CT:	Computed Tomography
CVPR:	Computer Vision and Pattern Recognition
DBN:	Deep Belief Network
DCE:	Dynamic Contrast Enhanced
DCIS:	Ductal Carcinoma In Situ
DDSM:	Digital Database for Screening Mammography
DenseNet:	Densely Connected Convolutional Network
DL:	Deep Learning
DM:	Digital Mammogram
DT:	Decision trees
DTT:	Discrete Tchebichef Transform
ELM:	Extreme Learning Machine
EME:	Effective Measure of Enhancement
FC:	Fully Connected
FN:	False Negatives
FNR:	False Negative Rate
FP:	False Positives
FPR:	False Positive Rate
FrCN:	Full Resolution Convolutional Network
GDFNN:	Generalized Dynamic Fuzzy Neural Networks
GLCM:	Gray-Level Co-occurrence Matrices
GLRLMs	Gray-Level Run-Length Matrices
GPUs:	Graphical Processing Units
IDC:	Invasive Ductal Carcinoma
ILC:	Invasive Lobular Carcinoma
JPEG:	Joint Photographic Experts Group

KeV:	Kiloelectron Volt
KNN:	K-Nearest Neighbors
LBP:	Local Binary Patterns
LDA:	Linear Discriminant Analysis
LE:	Low Energy
LoG:	Laplacian of Gaussian
LR:	Logistic Regression
mAs:	milliAmpere-seconds
MCC:	Matthews's Correlation Coefficient
MCs:	MicroCalcifications
MG:	MammoGraphy
MIAS:	Mammograms Image Analysis Society
ML:	Machine Learning
MLO:	MedioLateral Oblique
MLP:	Multi-Layer Perceptron
MRI:	Magnetic Resonance Imaging
NB:	Naive Bayes
OBL:	Opposition-Based Learning
PCA:	Principal Component Analysis
PET:	Positron Emission Tomography
PHIC:	Palestinian Health Information Center
PSNR:	Peak Signal to Noise Ratio
PSOWNN:	Particle Swarm Optimized Wavelet Neural Network
QT:	Quality Threshold
RBF:	Radial Basis Function
ReLU:	Rectified Linear Unit
ResNet:	Residual Network
RF:	Random Forests
RFE:	Recursive Feature Elimination
RNN:	Recurrent Neural Networks
ROC:	Receiver Operating Characteristic
ROI:	Region Of Interest
SCV:	Stratified Cross-Validation
SD-CNN:	Shallow-Deep Convolutional Neural Network
SLIC:	Simple Linear Iterative Clustering
SMO:	Sequential Minimal Optimization
SVM:	Support Vector Machines
TL:	Transfer learning
TN:	True Negatives
TNR:	True Negative Rate
TP:	True Positives
TPR:	True Positive Rate
U/S:	Ultrasound
VGGNet:	Visual Geometry Group Network
WHO:	World Health Organization
YOLO:	You Only Look Once

Chapter One : Introduction

In this chapter, we will provide an overview of several important components related to the study on breast cancer detection and classification using artificial intelligence from mammograms collected in the Palestine region. These components include the background, problem statement, justification, study objectives, hypothesis, and research question. By exploring each of these elements in detail, we aim to offer a comprehensive understanding of their significance within the context of the study.

1.1 Background

To facilitate the development of our research and the achievement of our desired goals, we will present a comprehensive series of topics. These topics will cover various aspects, starting with an understanding of breast cancer, including its incidence and mortality rates in Palestine. Additionally, we will explore the current methods used in the detection and classification of breast cancer. Lastly, we will provide a brief overview of artificial intelligence and its different forms.

1.1.1 Breast Cancer Statistics in Palestine

A breast tumor refers to an abnormal growth or mass of cells that develops within the breast tissue. These growths can be either non-cancerous (benign) or cancerous (malignant) and can occur in both men and women, although they are significantly more common in women (ACS, 2021). Benign breast tumors, such as fibroadenomas and cysts, do not spread to other parts of the body and are generally not life-threatening. While they may cause discomfort, pain, or changes in the appearance of the breast, benign tumors are typically not associated with an increased risk of breast cancer. In some cases, benign tumors may require monitoring or surgical removal if they continue to grow or cause significant symptoms. In contrast, malignant breast tumors are cancerous and have the potential to invade surrounding tissues and metastasize (spread) to other organs, a condition known as breast cancer (as shown in Figure 1). Malignant breast tumors come in various types, including ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC), among others. The specific type and characteristics of a malignant breast tumor, such as tumor grade, hormone receptor status, and genetic markers, significantly influence disease progression, treatment response, and the appropriate course of treatment (ACS, 2021).

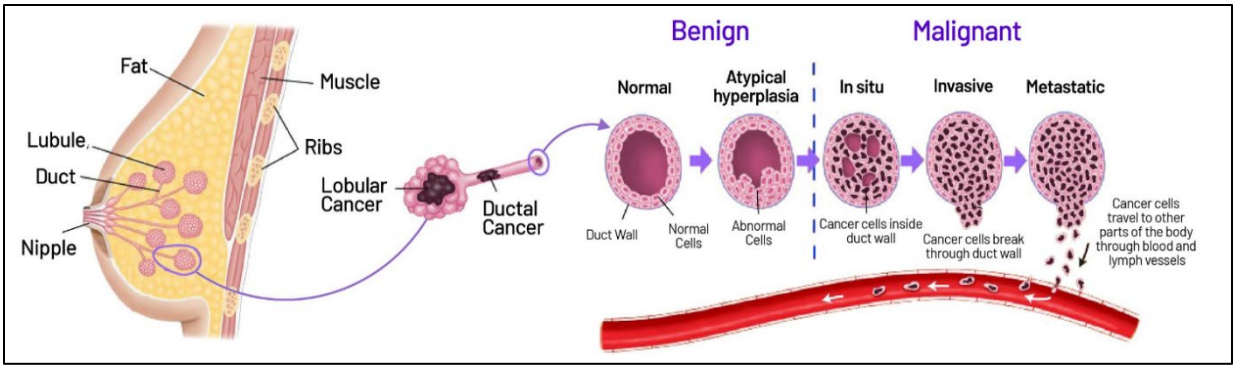

Figure 1: Progression of Breast Cancer (Saint John's Cancer Institute, 2024).

Breast cancer is a significant public health concern in Palestine, maintaining its position as the most prevalent cancer among the population, particularly among females. According to the Palestinian Health Information Center, in 2022, Palestine recorded a total of 934 new cases of breast cancer, translating to an incidence rate of 18.5 cases per 100,000 population. This figure places breast cancer as the most commonly diagnosed cancer in the country, surpassing colorectal cancer and lung cancer (PHIC, 2023). When examining the data at a regional level, the West Bank reported 540 new breast cancer cases, constituting 15.8% of all newly registered cancer cases and resulting in an incidence rate of 18.8 cases per 100,000 population, as shown in Figure (2 & 3). Similarly, the Gaza Strip witnessed 394 new cases of breast cancer, accounting for 19.2% of all new cancer cases and yielding an incidence rate of 18.2 cases per 100,000 people, as shown in Figure 2 (PHIC, 2023).

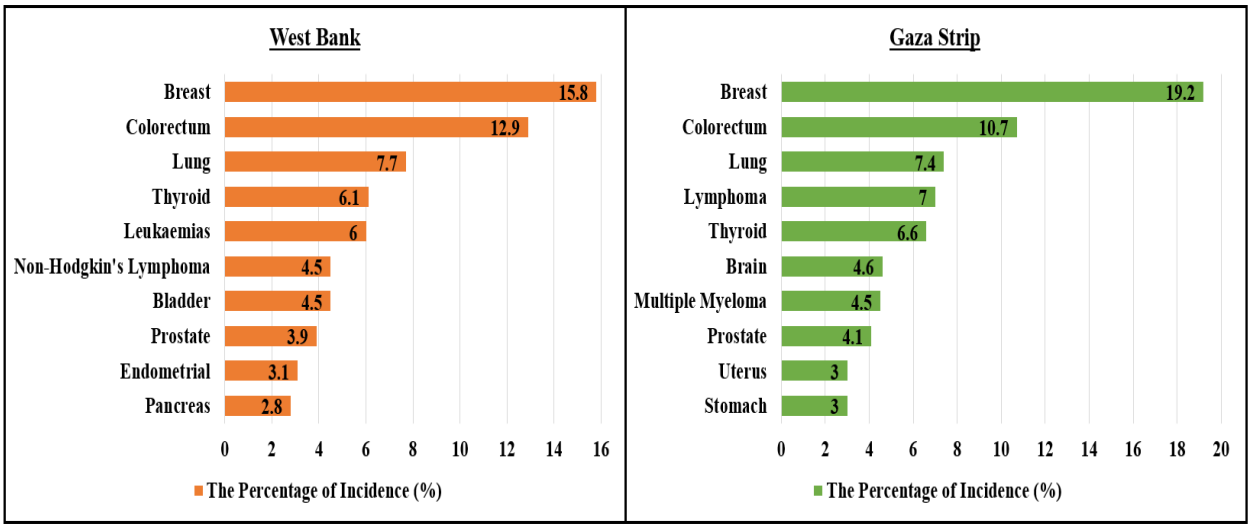

Figure 2: Distribution of Percentage of Top Ten Reported Cancers in all population, Palestine 2022 (PHIC, 2023).

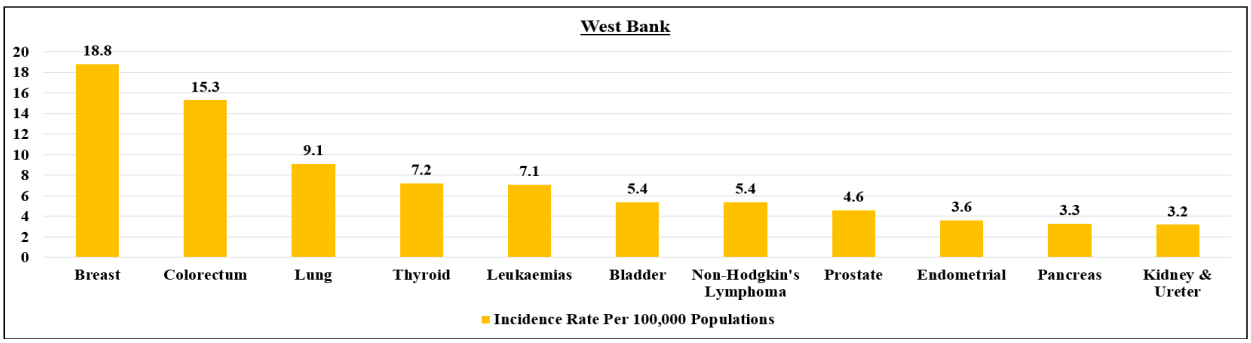

Figure 3: Incidence Rate of Top Ten Reported Cancers per 100,000 of population, West Bank, Palestine 2022 (PHIC, 2023).

Among the Palestinian population, breast cancer maintains its position as the most widespread cancer, particularly among females, with an incidence rate of 38 cases per 100,000 population. This high incidence rate among females is a cause for concern and highlights the need for targeted awareness campaigns and screening programs tailored to the Palestinian context. However, for the male population, colorectal cancer takes the lead as the most frequently diagnosed cancer type, while leukemia emerges as the predominant form of cancer among the pediatric population, as shown in Figure 4 (PHIC, 2023). Despite its high incidence rate, especially among females, breast cancer secures the third position among the ten most fatal types of cancer in the West Bank, representing an 11.7% mortality rate, as shown in Figure 5. The recorded data in Table 1 indicates that there were a total of 144 deaths due to breast cancer in 2022, with 8 cases in males and 136 cases in females. This stark disparity in mortality between genders highlights that the majority of breast cancer fatalities occur in the female population (PHIC, 2023).

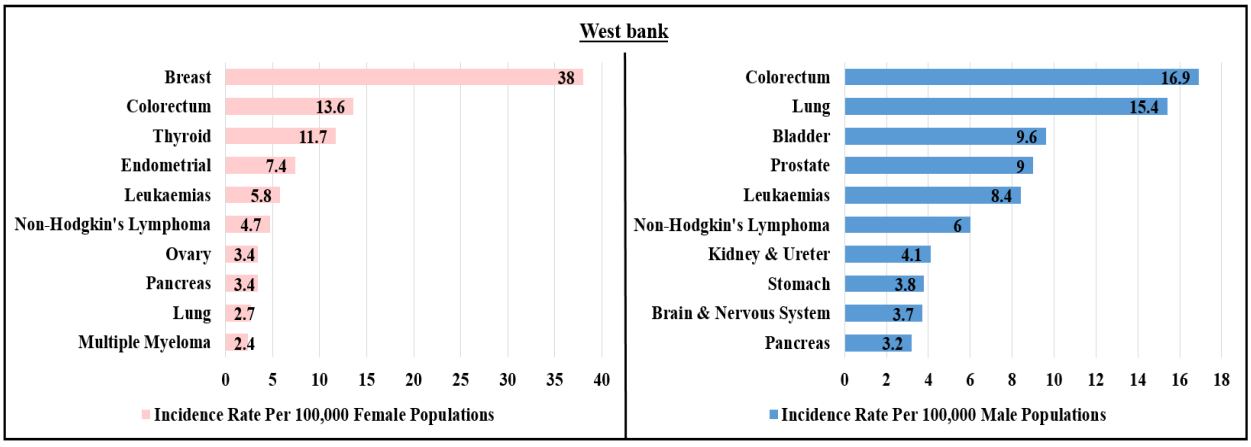

Figure 4: Incidence rates per 100,000 population for the top ten reported cancer types among males and females in the West Bank, Palestine, in 2022 (PHIC, 2023).

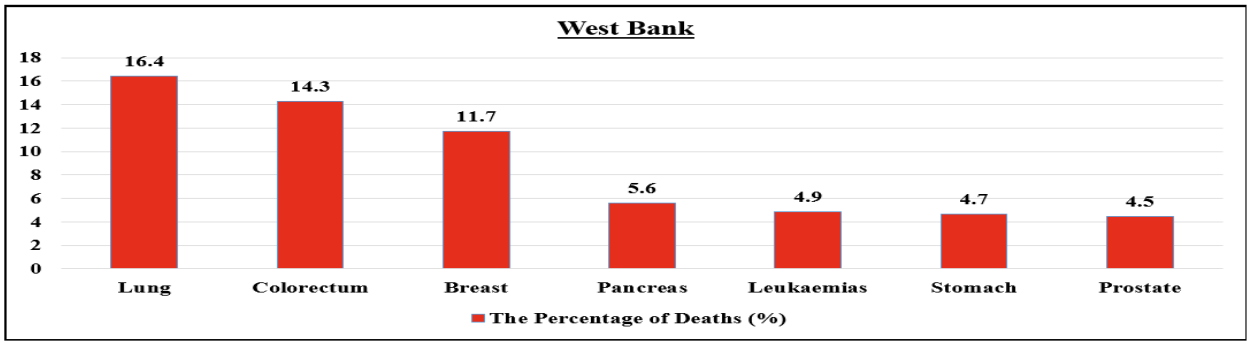

Figure 5: Proportional Distribution of the most Reported Cancer Deaths of all Reported Cancer Deaths, West Bank, 2022 (PHIC, 2023).

Table 1: Distribution of Reported Cancer Deaths by Site & Sex, West Bank 2022 (PHIC, 2023).

Site	International Statistical Classification of Diseases and Related Health Problems (ICD)	West Bank			
		Male	Female	Total	Mortality Rate (%)
Lung	C34	184	18	202	16.4%
Colorectal	C18-C20	102	74	176	14.3%
Breast	C50	8	136	144	11.7%
Pancreas	C25	37	32	69	5.6%
Leukemia	C91-C95	37	24	61	4.9%
Stomach	C16	35	23	58	4.7%
Prostate	C61	55	0	55	4.5%

The global landscape of breast cancer highlights significant disparities in incidence and mortality rates between high-income and low-to-middle-income countries. Generally, high-income countries tend to have higher rates of breast cancer incidence, attributed to factors such as westernized lifestyles, delayed childbearing, and lower breastfeeding rates. However, these countries often have better access to healthcare, including early detection through screening programs and advanced treatment options, contributing to lower mortality rates. In contrast, low and middle-income countries often experience higher mortality rates due to challenges in accessing quality healthcare, limited availability of mammography screening, late-stage diagnosis, and suboptimal treatment access and affordability (Francies et al., 2020).

A 2022 study conducted by the World Health Organization (WHO) provides valuable insights into the specific breast cancer rates in the Western Asia region, which includes Palestine. The study found that the age-standardized incidence and mortality rates in this region were 45.4 and 15.1 per 100,000 females, respectively. Notably, Palestine, despite being classified as a low-income country, reports even higher incidence and mortality rates of 46.3 and 19.7 per 100,000 females, respectively (as shown in Figure 6) (WHO, 2022b). This data underscores the urgent need to address the public health impact of breast cancer in Palestine, as it is the most common cancer among females, accounting for 15.8% of all newly registered cancer cases in the West Bank and 19.2% in the Gaza Strip, and ranks third in terms of mortality (11.7%) after lung and colorectal cancer (PHIC, 2023).

These findings highlight the necessity for targeted interventions to improve early detection, enhance access to quality treatment, and strengthen the overall breast cancer care infrastructure in the region. Efforts should be made to increase awareness about breast cancer risk factors, promote regular screening through mammography, and ensure timely access to diagnostic and treatment services, particularly for women in low-resource settings. Additionally, investing in research to understand the unique socio-cultural, environmental, and genetic factors contributing to the high incidence and mortality rates in Palestine could inform the development of tailored prevention and control strategies (WHO, 2022b).

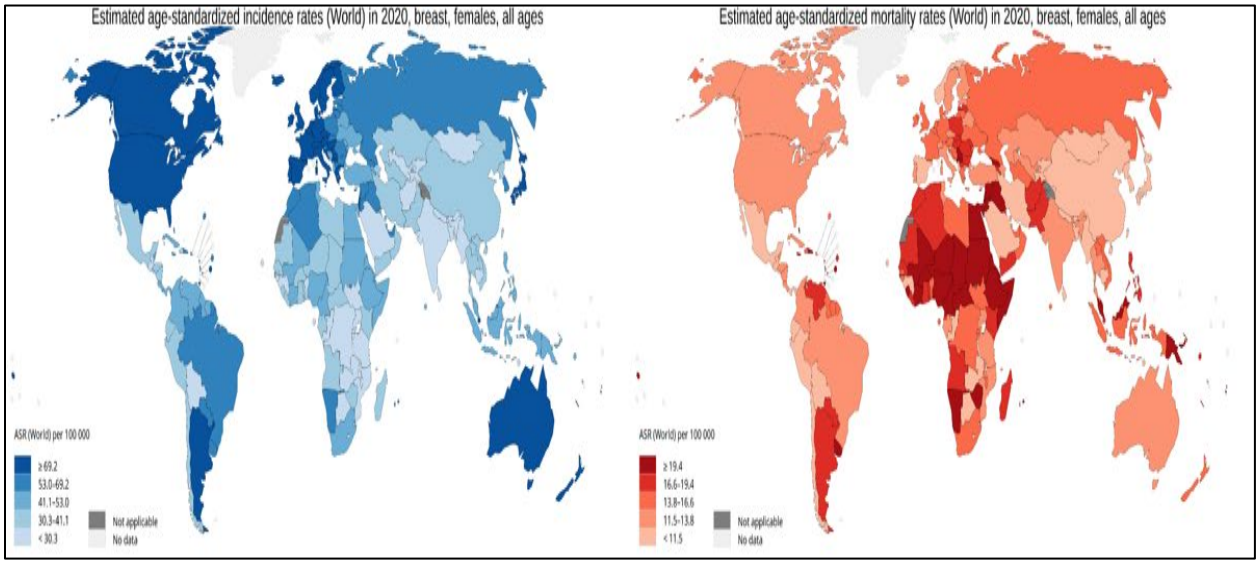

Figure 6: A global map representing the incidence and mortality rates of breast cancer, World Health Organization in 2022 (WHO, 2022a).

1.1.2 Detection and Classification of Breast Cancer

Early detection of breast cancer is crucial for successful treatment and improved patient outcomes, as it increases the chances of detecting the disease at an early stage when it is more localized and easier to treat (Kashyap et al., 2022). . Regular breast self-examinations (BSE) and routine breast cancer screenings, such as mammography, are effective methods for detecting breast cancer at an early stage (Siegel et al., 2018). A meta-analysis published in the International Journal of Cancer found that women who practiced BSE had a 25% lower risk of presenting with late-stage breast cancer compared to those who did not practice BSE. This finding highlights the importance of encouraging women to perform regular BSEs as a complementary method to clinical examinations and mammography screening.

Mammography remains the gold standard technique for breast cancer screening and diagnosis. This imaging method utilizes low-energy X-rays, typically in the range of 20-30 keV, to identify any abnormalities or suspicious areas within the breast, such as masses or microcalcifications, as shown in Fig (7) (Giampietro et al., 2020). The key advantage of mammography is its ability to detect breast cancer at an early stage, often before any visible symptoms appear. Early detection is crucial, as it significantly increases the chances of successful treatment and improves patient survival rates (Heywang-Köbrunner et al., 2011). According to a study published in the Journal of Medical Screening, implementing organized mammography screening programs can reduce breast cancer mortality by approximately 20-25% in women aged 50-69 years. The study further revealed that for every 1,000 women screened biennially (every two years) over a 20-year period, around 7-9 breast cancer deaths could potentially be prevented.

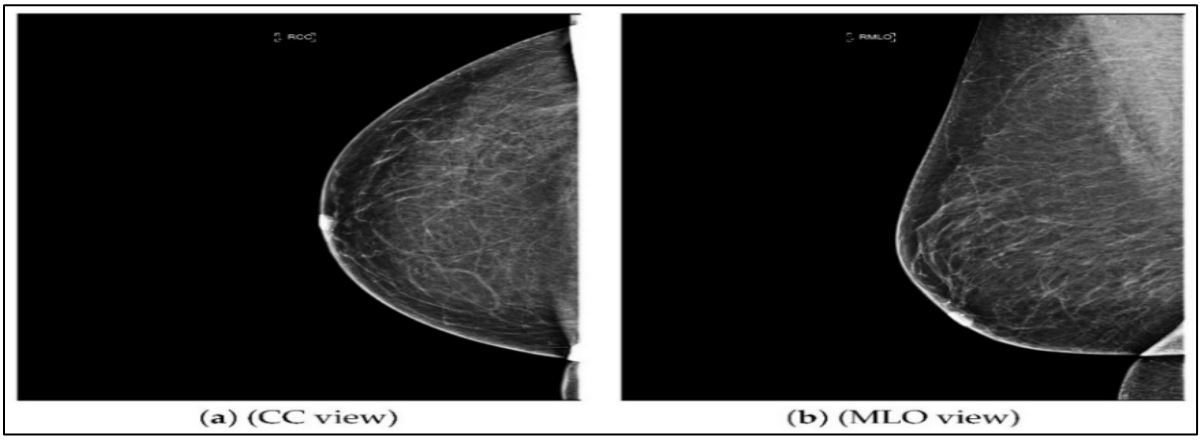

Figure 7: Breast mammograms (MD) taken from two different positions (Justaniah et al., 2022).

However, mammography has limitations. Its sensitivity, or the true positive rate, is approximately 85% (Zhu et al., 2023). This sensitivity can decrease to around 50% in middle-aged individuals with denser breast tissue, making it more challenging to distinguish between malignant and benign cases (Zhao et al., 2015). To address this limitation, healthcare professionals may employ additional diagnostic techniques, such as clinical breast examinations (CBE), ultrasound-guided tru-cut biopsy, ultrasound, or magnetic resonance imaging (MRI), to complement mammography and improve the overall accuracy of breast cancer detection (Karellas & Vedantham, 2008), as shown in Figure 8 and Table 2. A study in the Journal of Medical Screening reported that CBE combined with mammography screening could increase the detection rate of early-stage breast cancers by up to 10%.

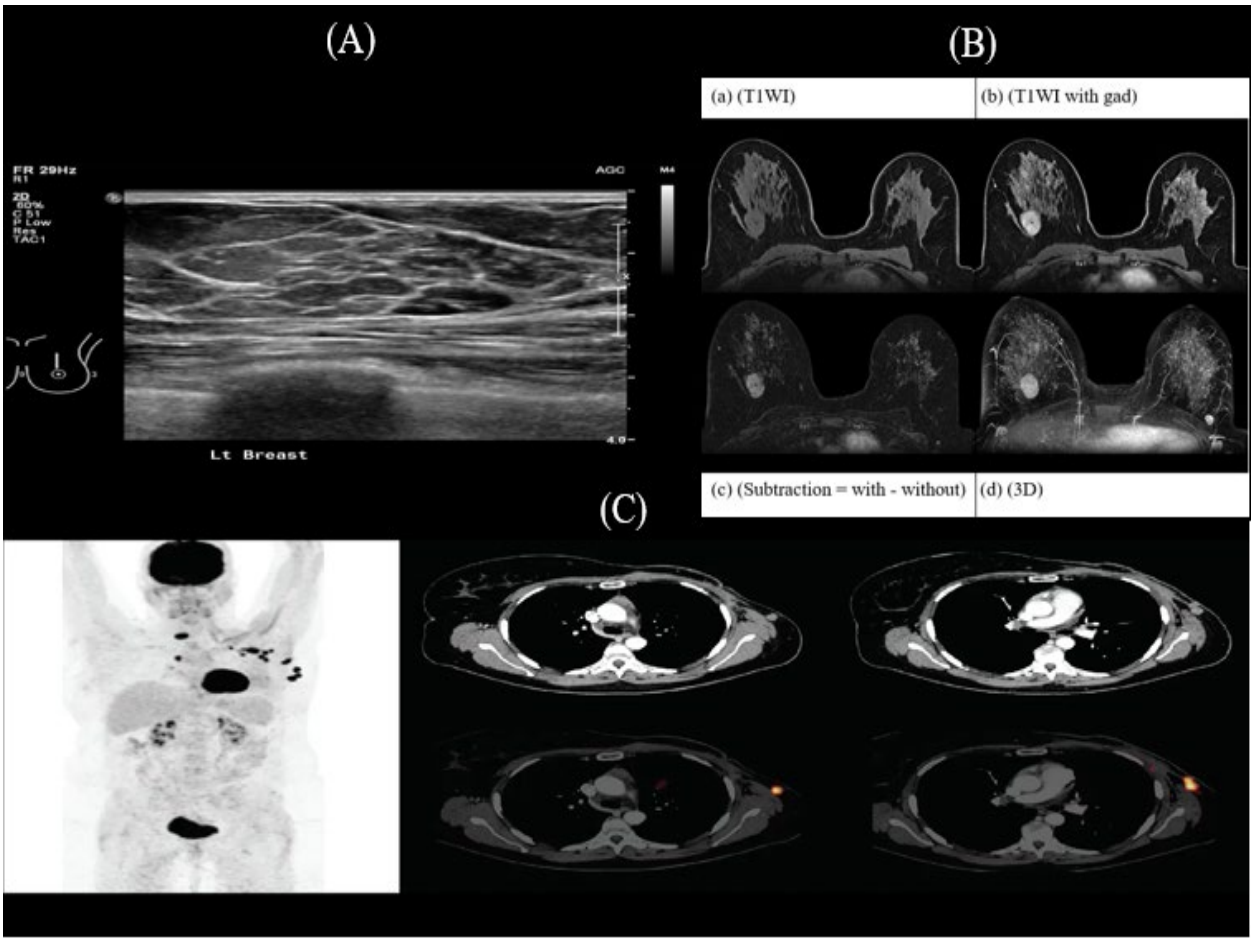

Figure 8: Imaging Methods for Breast Cancer Detection: Ultrasound (A), Magnetic Resonance Imaging (MRI) (B), Computed Tomography, and PET Scan (C).

Table 2: Comparison of different types of breast imaging (Zhu et al., 2023).

Modality	Sensitivity	Tumor size	Advantages	Disadvantages
MG	85%	≤2 cm	-Detects the early-stage breast cancer. -Improved image resolution. -Widely available.	-Use of ionizing radiation. -Low specificity may cause unnecessary biopsies. -Limited sensitivity in dense breast tissue.
US	82%	2 cm	-No radiation. -Suitable for pregnant patients, dense breasts and implant imaging. -Safe and low cost compared with MG.	-High requirement for operator. -Operator-dependent, limited specificity.
MRI	95%	≤2 cm	-Suitable for high-risk patients. -High sensitivity. -Images small details of soft tissues.	-Low specificity. -High cost compared with MG and US.

Moreover, ongoing research is exploring the potential of emerging technologies, like positron emission tomography (PET) and computed tomography (CT), to enhance the early diagnosis and monitoring of breast cancer (Anandhamala, 2018). According to a study published in the Journal of the American Medical Association, the use of advanced breast imaging combined with conventional mammography increased the detection rate of invasive breast cancers by 41% compared to mammography alone. These findings highlight the potential benefits of integrating multiple screening and diagnostic techniques to enhance the early detection of breast cancer, thereby improving treatment outcomes and reducing mortality rates. Despite its limitations, mammography remains the cornerstone of breast cancer screening, offering a reliable and widely available method for the early identification of this disease. Continued advancements in imaging technology and the integration of complementary diagnostic approaches can further improve the accuracy and effectiveness of breast cancer detection, ultimately leading to better patient outcomes.

Accurate and high-quality imaging is crucial for the effective detection and management of breast cancer. Mammography, ultrasound, and magnetic resonance imaging (MRI) are the primary imaging modalities used, and the quality of the images obtained through these techniques is of paramount importance. Factors such as image resolution, positioning, compression technique, and equipment calibration can all impact the overall quality of the images. Advancements in technology, including digital mammography, tomosynthesis (3D mammography), color doppler, and dynamic contrast-enhanced MRI (DCE-MRI), have significantly improved image quality and enhanced the detection capabilities of these imaging techniques (Azhdeh et al., 2021).

Radiologists play a vital role in the interpretation, analysis, and classification of breast imaging, contributing their expertise and experience to the accurate diagnosis and management of breast cancer. They carefully examine images from various modalities, including mammography, ultrasound, and MRI, looking for subtle changes or abnormalities that may indicate the presence of breast cancer, such as masses, calcifications, architectural distortions, or suspicious enhancement patterns. Through their specialized training and extensive experience, radiologists can differentiate between benign and malignant findings, determine the need for further diagnostic tests or interventions, and provide accurate and timely diagnoses (Yeh et al., 2013). However, the methods used by radiologists, which heavily rely on human visual perception and cognitive abilities, can be subject to variability and potential errors, impacting the accuracy of breast cancer diagnosis. Studies have shown that the false-negative rate in mammography interpretation can range from 10% to 30%, meaning that a substantial proportion of cancers may be missed during initial screening (Ekpo et al., 2018). This highlights the importance of developing standardized classification systems and implementing quality assurance measures to ensure consistent and reliable interpretation of breast imaging findings.

One of the key tools used by radiologists for breast tumor classification is the Breast Imaging Reporting and Data System (BI-RADS), a standardized system developed by the American College of Radiology (ACR). The BI-RADS system categorizes breast lesions into different levels of suspicion, ranging from BI-RADS 0 (incomplete assessment) to BI-RADS 6 (known malignancy), based on various features such as shape, margins, density, and the presence of calcifications, as shown in Table 3 (ACR, 2023). These BI-RADS categories guide further management decisions, such as the need for additional diagnostic tests, biopsy, or close

monitoring, and help ensure consistent communication and understanding among healthcare professionals involved in the patient's care. The BI-RADS system is not only used for mammography but is also applied to breast ultrasound and MRI findings (ACR, 2023).

Table 3: Breast Imaging Reporting and Data System (BI-RADS) categories (Smithuis, 2014).

Category		Management	Likelihood of cancer
0	Need additional imaging or prior examinations	Recall for additional imaging and/or await prior examinations	n/a
1	Negative	Routine screening	Essentially 0%
2	Benign	Routine screening	Essentially 0%
3	Probably Benign	Short interval-follow-up (6 month) or continued	> 0% but ≤ 2%
4	Suspicious	Tissue diagnosis	4a. low suspicion for malignancy (> 2% to ≤ 10%)
			4b. moderate suspicion for malignancy (> 10% to ≤ 50%)
			4c. high suspicion for malignancy (> 50% to < 95%)
5	Highly suggestive of malignancy	Tissue diagnosis	≥ 95%
6	Known biopsy-proven	Surgical excision when clinical appropriate	n/a

Radiologists evaluate various features of the lesions, such as shape, margins, echogenicity, and vascularity for ultrasound, and lesion morphology, enhancement kinetics, and enhancement patterns for MRI, to assign a BI-RADS category (Abdulloh & Ni'mah, 2023). Studies have shown that the positive predictive value (PPV) for malignancy increases with higher BI-RADS categories across different imaging modalities (Heinig et al., 2008; Strigel et al., 2017; Strobel et al., 2015). For example, a study published in the journal Radiology in 2015 analyzed the accuracy of BI-RADS categories in predicting breast cancer based on mammography findings. The study involved 1,051 lesions and found that the PPV for malignancy increased from 23.8% for BI-RADS category

4 (suspicious abnormality) to 97.4% for category 5 (highly suggestive of malignancy) (Freer et al., 2015).

In addition to the BI-RADS system, radiologists may employ other advanced techniques, such as Kinetic Curve Analysis, to further characterize breast lesions detected on MRI. This technique involves analyzing the enhancement patterns of breast lesions over time, assessing the kinetics of contrast uptake and washout (Yim et al., 2016). Different enhancement patterns, such as persistent, plateau, or washout, can provide valuable information about the likelihood of malignancy, as shown in Figure 9. A study published in the journal Breast Cancer Research and Treatment in 2018 investigated the accuracy of BI-RADS categories in predicting breast cancer based on MRI findings. The study included 1,105 breast lesions and found that the PPV for malignancy increased with higher BI-RADS categories, ranging from 4.3% for category 3 to 92.6% for category 5 (DeMartini et al., 2011).

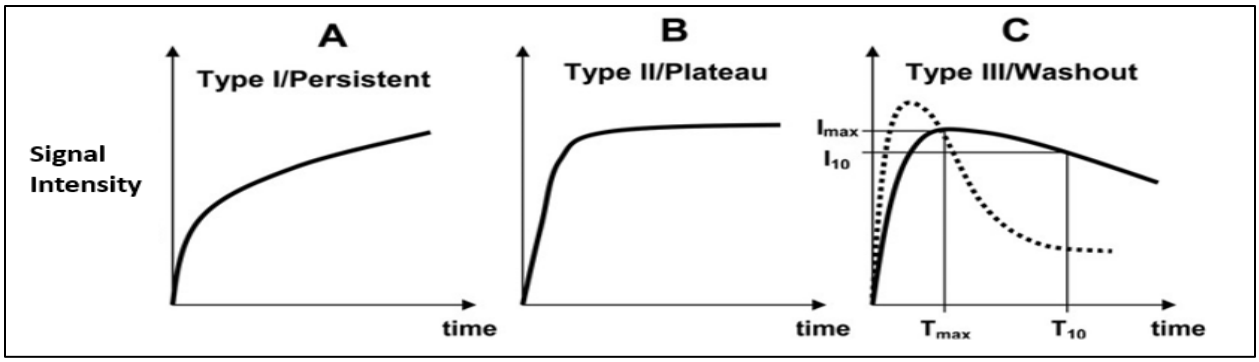

Figure 9: Graphs depicting three patterns of Kinetic curves typically seen in breast lesions, as intensity enhancement as a function of time. A) Type I – (persistent enhancing), B) Type II – (plateau) and C) Type III – (washout) (Craciunescu et al., 2009).

Radiologists work closely with other healthcare professionals, including breast surgeons, oncologists, and pathologists, to ensure a comprehensive and coordinated approach to patient care. They collaborate with the healthcare team to discuss cases, review imaging findings, and develop personalized treatment plans for patients. Additionally, radiologists are responsible for monitoring treatment response and surveillance imaging to detect any potential recurrence or new developments (Kuhl, 2015). However, it is important to note that access to high-quality healthcare services plays a pivotal role in addressing the impact of breast cancer. Ensuring equitable access to timely and appropriate treatment within a well-equipped healthcare infrastructure is crucial for effective management of the disease. Disparities in access to healthcare services can have

significant implications for breast cancer outcomes, underscoring the importance of addressing these inequities (Horton et al., 2020). A study published in the Journal of the National Cancer Institute found that women who received regular mammography screening had a 47% lower risk of dying from breast cancer compared to those who did not receive regular screening. This highlights the critical role of screening programs and early detection in improving breast cancer survival rates.

1.1.3 The Implementation of Artificial Intelligence (AI) in Breast Imaging

Artificial intelligence (AI) is a multidisciplinary field that combines computer science, mathematics, and other disciplines to develop intelligent systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, problem-solving, and decision-making (Boden, 1996). The advancement of AI is driven by various techniques, with machine learning being a prominent one. Machine learning involves training algorithms on large datasets to recognize patterns and make informed decisions. As the AI system is exposed to more data and experiences, its performance improves over time, allowing it to become increasingly adept at the tasks it is designed to perform (Michalski et al., 2013). A subset of machine learning, known as deep learning, utilizes artificial neural networks inspired by the human brain's structure and function. These neural networks are composed of multiple interconnected layers that can learn to extract and recognize intricate patterns and representations from raw data, such as medical images (Ker et al., 2018). Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in various computer vision tasks, including medical image analysis, due to their ability to automatically learn and extract relevant features from the images without the need for manual feature engineering (Yasaka & Abe, 2018).

The integration of AI technology in medical imaging brings several significant benefits. Firstly, AI algorithms can improve diagnostic accuracy by analyzing medical images with high precision and consistency. They can detect subtle abnormalities, lesions, or patterns that may be missed by human observers, leading to earlier detection and improved patient outcomes (Ito et al., 2022). Secondly, AI enhances efficiency by processing large volumes of medical images quickly, reducing the time required for diagnosis. This can help radiologists prioritize cases, leading to faster treatment decisions and reduced waiting times for patients (Ahuja, 2019). AI also provides valuable decision support to radiologists. By highlighting areas of concern, suggesting potential

diagnoses, or quantifying the likelihood of malignancy, AI algorithms can assist radiologists in making more informed decisions and improving diagnostic accuracy (Ahuja, 2019). Additionally, AI helps standardize the interpretation and classification of medical images. By following predefined algorithms and decision rules, AI systems can reduce inter-observer variability and ensure consistent and reliable results across different healthcare settings (Hah & Goldin, 2021).

In recent years, AI has revolutionized the field of medical imaging, particularly in diagnosing and classifying medical images used for breast cancer detection. AI algorithms, specifically machine learning and deep learning models, have been designed and trained on vast datasets of medical images to analyze and interpret these images with a high degree of accuracy. These AI models leverage advanced pattern recognition, feature extraction, and data analysis capabilities to provide valuable support to radiologists in the early identification and categorization of breast lesions, abnormalities, and potential malignancies. By automating the detection and analysis of subtle patterns and features that might otherwise be missed by human observers, AI systems can aid in early intervention and treatment, while also reducing the occurrence of false positives and unnecessary procedures, ultimately enhancing patient care and outcomes (Coppola et al., 2021).

In the context of breast cancer imaging, AI algorithms have shown great promise in improving the efficiency and precision of diagnosis. Computer-aided detection (CAD) systems, which utilize AI techniques such as machine learning and deep learning, have been developed to assist radiologists in identifying suspicious lesions or abnormalities on mammograms, ultrasound, or MRI images (Jalalian et al., 2013; Tang et al., 2009). These CAD systems can analyze the images and highlight areas of concern, such as masses, calcifications, architectural distortions, or suspicious enhancement patterns that may require further evaluation. Studies have shown that CAD systems can improve the sensitivity of radiologists in detecting breast cancer by up to 20%, potentially reducing the number of missed cases (Pacilè et al., 2020). However, it is important to note that CAD systems are intended to be used as a second reader, complementing the radiologist's interpretation rather than replacing it entirely. Beyond CAD, AI algorithms are also being developed for automated breast lesion classification and risk assessment. These algorithms can analyze various features of breast lesions, such as shape, margins, texture, and enhancement patterns, and classify them into different categories based on their likelihood of malignancy (Pacilè et al., 2020; Sadoughi et al., 2018; Sheth & Giger, 2020). Deep learning models, particularly

CNNs, have shown excellent performance in breast lesion classification tasks, achieving accuracy levels comparable to or even surpassing those of experienced radiologists (Murtaza et al., 2020). Furthermore, AI can play a role in risk stratification and personalized screening recommendations. By analyzing a combination of imaging features, clinical data, and other risk factors, AI algorithms can identify individuals at higher risk for developing breast cancer and suggest personalized screening intervals or additional imaging tests (Sheth & Giger, 2020). This can improve the overall effectiveness of breast cancer screening programs and ensure that high-risk individuals receive appropriate and timely diagnostic evaluations.

It is important to note that the successful implementation of AI in breast cancer imaging requires high-quality training data, rigorous validation, and careful integration into clinical workflows. Radiologists' expertise and clinical judgment remain crucial in interpreting AI-generated results, considering the clinical context, and making the final diagnosis and treatment decisions. The collaboration between AI and radiologists can lead to more accurate and timely diagnoses, ultimately improving patient care and outcomes (Rubin, 2019). However, there are also challenges and ethical considerations that must be addressed in the development and deployment of AI systems in healthcare. These include ensuring data privacy and security, mitigating potential algorithmic biases, achieving transparency and explain ability in AI decision-making, and addressing issues related to liability and accountability (Alonso & Siracuse, 2023; Hlávka, 2020). Ongoing research and collaboration between medical professionals, computer scientists, and AI experts are essential to further develop and refine AI-based solutions for breast cancer imaging. This includes exploring new AI techniques, improving model performance and robustness, and conducting large-scale clinical trials to evaluate the real-world impact of AI on patient outcomes. Additionally, addressing the ethical and regulatory challenges surrounding AI in healthcare will be crucial for the responsible and equitable adoption of these technologies (Shah et al., 2022).

1.2 Problem Statement

Women are the backbone of society, playing a crucial role in nurturing families, bringing new life into the world, and shaping the future through their influence on the next generation. However, the impact of breast cancer on women is significant, with the disease claiming many lives within this demographic, particularly in Palestine. Therefore, the development of an artificial intelligence system designed to accurately detect and classify breast cancer in its early stages is of utmost

importance to alleviate the suffering of this population. Mammography is a widely used screening technique for breast cancer detection, but the interpretation of mammogram images can be challenging and subjective, leading to potential misdiagnoses or missed cases as noted in the previous background. Traditional computer-aided detection (CAD) systems for mammogram analysis often rely on hand-crafted features and conventional machine learning algorithms, which may not fully capture the complex patterns and subtle differences present in mammogram images. Moreover, these systems typically treat the feature extraction and classification stages as separate processes, potentially limiting their overall performance. Recent advances in deep learning and machine learning techniques have shown promising results in various medical image analysis tasks, including breast cancer detection and classification from mammogram images. Deep learning models, such as convolutional neural networks (CNNs), have demonstrated remarkable performance in automatically learning discriminative features from raw image data. However, these models may still benefit from the incorporation of traditional machine learning classifiers, which can leverage the learned features and provide robust classification capabilities.

To address these challenges, this study investigates the use of hybrid models that combine deep learning for feature extraction and machine learning for classification in the context of breast cancer detection and diagnosis from mammogram images. Specifically, we explore the performance of hybrid models that combine two popular CNN architectures, VGG16 and DenseNet121, for feature extraction with four widely used machine learning classifiers: random forest, gradient boosting, support vector machine, and logistic regression. Furthermore, we assess the impact of image enhancement techniques, such as morphological erosion, Contrast-Limited Adaptive Histogram Equalization (CLAHE), Laplacian of Gaussian (LoG) edge enhancement and Unsharp Masking, on the performance of these hybrid models. These techniques aim to improve the visual quality and contrast of mammogram images, potentially enhancing the feature extraction and classification processes. By evaluating the performance of these hybrid models on a dataset of mammogram images collected from a center in Palestine, we aim to identify the most effective combinations of deep learning and machine learning techniques for accurate breast cancer detection and classification. Additionally, we investigate the generalization capabilities of the hybrid models by applying them to new, unseen mammogram images and comparing their performance to the initial training and validation sets. Ultimately, this study seeks to contribute to

the development of reliable and efficient CAD systems for breast cancer detection and diagnosis, potentially improving patient outcomes and reducing the burden of this devastating disease.

1.3 Research Objectives

The overarching goal of this research is to develop and evaluate effective hybrid models that combine deep learning and machine learning techniques for the accurate detection and classification of breast cancer from mammogram images. To achieve this goal, several specific research objectives have been formulated:

1. To develop and evaluate the performance of eight hybrid models that combine deep learning feature extraction techniques (VGG16 and DenseNet121) with machine learning classifiers (Random Forest, Gradient Boosting, Support Vector Machine, and Logistic Regression) for detecting and classifying benign and malignant breast cancer from mammogram images.
2. To investigate the impact of image enhancement techniques, specifically morphological erosion, Contrast-Limited Adaptive Histogram Equalization (CLAHE), Laplacian of Gaussian (LoG) edge enhancement and Unsharp Masking, on the performance of the hybrid models in terms of accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC), and the time spent in the prediction process..
3. To assess the generalization capabilities of the hybrid models by applying them to new, unseen mammogram images and comparing their performance metrics (accuracy, precision, recall, F1-score, AUC and TIME) to those obtained on the initial training and validation sets.
4. To identify the most effective combination(s) of deep learning feature extractor and machine learning classifier among the eight hybrid models, based on their performance in breast cancer detection and classification from mammogram images.

1.4 Research Questions

This study aims to investigate the performance of hybrid models that combine deep learning and machine learning techniques for breast cancer detection and classification from mammogram images. Specifically, we seek to answer several research questions related to the effectiveness of these hybrid models, the impact of image enhancement techniques, and the generalization

capabilities of the models when applied to new, unseen data. The research questions guiding this study are as follows:

Q1: How do different hybrid models combining deep learning feature extraction (VGG16 and DenseNet121) and machine learning classifiers (Random Forest, Gradient Boosting, Support Vector Machine, and Logistic Regression) perform in detecting and classifying benign and malignant breast cancer from mammogram images specifically in the context of healthcare in Palestine?

Q2: What are the effects of image enhancement techniques, such as morphological erosion, Contrast-Limited Adaptive Histogram Equalization (CLAHE), Laplacian of Gaussian (LoG) edge enhancement and Unsharp Masking, on the performance of the hybrid models in terms of accuracy, precision, recall, F1-score, and AUC?

Q3: How do the hybrid models perform when applied to new, unseen mammogram images, and how does their performance compare to the initial training and validation sets?

Q4: Which hybrid model(s) achieve the best overall performance in terms of accuracy, precision, recall, F1-score, and AUC for breast cancer detection and classification from mammogram images?

Q5: How do the computational times (TIME) of the hybrid models compare, and what are the trade-offs between model performance and computational efficiency?

These research questions can help guide analysis and discussion of the hybrid models' performance, the effects of image enhancement techniques, and the potential for practical application in breast cancer detection and classification using mammogram images.

1.5 Research Justifications

Breast cancer is a significant public health concern worldwide, and early detection through screening mammograms is crucial for improving treatment outcomes and saving lives. The motivation behind this research stems from the pressing need to improve the accuracy and efficiency of breast cancer detection and diagnosis from mammogram images. However, the interpretation of mammograms can be challenging, even for experienced radiologists, due to the complexity and subtlety of the images. Misinterpretations can lead to missed diagnoses or false

positives, which can have serious consequences for patients. Artificial intelligence (AI) techniques, particularly machine learning algorithms and deep convolutional neural networks, have shown great promise in analyzing medical images and assisting in disease detection and classification. By training these algorithms on large datasets of mammograms, they can learn to recognize patterns and features associated with breast cancer, potentially improving the accuracy and consistency of diagnosis.

In Palestine, access to advanced diagnostic tools and specialized medical expertise may be limited, particularly in remote or underserved areas. Implementing AI-based systems for mammogram analysis could help bridge this gap and provide a valuable screening tool, enabling earlier detection and intervention for breast cancer cases. Furthermore, the mammogram images collected from the center in Palestine represent a unique dataset that could contribute to the development and validation of AI models for breast cancer detection. Different populations may exhibit variations in breast tissue characteristics, breast density patterns, and the presentation of breast cancer lesions. By training AI models on diverse datasets, including those from Palestinian women, the algorithms can become more robust and generalizable, enhancing their performance across different populations and settings. Successful implementation of AI-based mammogram analysis in Palestine could not only improve breast cancer screening and early detection efforts but also serve as a model for other regions facing similar challenges. The research findings could inform the development of cost-effective and accessible AI-based screening solutions, potentially saving lives and improving healthcare outcomes in resource-limited settings.

1.6 Research Hypotheses

This study aims to investigate the effectiveness of hybrid models that combine deep learning and machine learning techniques for breast cancer detection and classification from mammogram images. Several research hypotheses have been formulated to guide the analysis and interpretation of the results, as well as evaluate the performance and robustness of the hybrid models. The primary hypothesis is that the hybrid models, which utilize deep learning feature extraction techniques (VGG16 and DenseNet121) and machine learning classifiers (Random Forest, Gradient Boosting, Support Vector Machine, and Logistic Regression), will outperform traditional methods in accurately detecting and classifying benign and malignant breast cancer from mammogram images. These hybrid models are expected to achieve superior performance in terms of evaluation

metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Additionally, it is hypothesized that applying image enhancement techniques, specifically morphological erosion, Contrast-Limited Adaptive Histogram Equalization (CLAHE), Laplacian of Gaussian (LoG) edge enhancement, and Unsharp Masking, will further improve the performance of the hybrid models by enhancing the visual quality and contrast of the mammogram images.

Another key hypothesis is that the hybrid models will demonstrate robust generalization capabilities when applied to new, unseen mammogram images. The performance metrics achieved by the hybrid models on these new images are expected to be comparable to those obtained on the initial training and validation sets, highlighting their ability to generalize well to previously unseen data. Additionally, it is hypothesized that certain combinations of deep learning feature extractors and machine learning classifiers will exhibit superior performance, identifying the most effective combinations for this task. The computational time required for training and inference of the hybrid models is also anticipated to vary, with certain models offering a better trade-off between performance and computational efficiency. Furthermore, it is hypothesized that incorporating ensemble techniques or hyperparameter tuning in the hybrid models will lead to higher performance compared to baseline models. Ensemble methods, such as bagging or boosting, and the optimization of hyperparameters for both the deep learning and machine learning components are expected to further enhance the performance of the hybrid models by leveraging the strengths of multiple models or optimizing their configurations.

Finally, it is hypothesized that the hybrid models proposed in this study will outperform traditional computer-aided detection (CAD) systems or single-model approaches (either deep learning or machine learning alone) in terms of overall performance and robustness for breast cancer detection and classification from mammogram images. The combination of deep learning and machine learning techniques is expected to provide complementary advantages, resulting in superior performance compared to methods that rely solely on one type of approach.

Chapter Two: Literature Review

This chapter delves into the advancements made in the application of artificial intelligence (AI) techniques for detecting and classifying breast cancers in mammogram images. The use of both machine learning and deep learning approaches has transformed this field, offering significant potential for assisting radiologists in accurately diagnosing breast abnormalities. However, it is crucial to optimize each method appropriately for the specific task and dataset to achieve the highest diagnostic performance. Consequently, this chapter examines the development and application of computational methods, such as comparing different neural network architectures or machine learning algorithms, fine-tuning hyperparameters, incorporating additional data sources, and implementing preprocessing and segmentation techniques. By comprehensively understanding the workflows of machine learning and deep learning used in this context, valuable insights can be gained for the design of future systems. To begin, the foundations of machine learning and the latest developments in this field relevant to the analysis of medical imaging will be discussed.

2.1 Machine Learning Applications in Breast Cancer Detection and Classification

The integration of artificial intelligence (AI) into the field of medical imaging is heavily reliant on the availability of diverse and extensive datasets. These datasets encompass a wide range of information, including various types of medical images such as X-rays, ultrasounds, and MRIs, as well as patient records and clinical data sourced from various healthcare institutions like hospitals, research facilities, and medical imaging centers (Le et al., 2019). Access to these comprehensive datasets is crucial for training AI algorithms to effectively recognize patterns, make accurate predictions, and aid in medical decision-making processes. AI methods employed in medical imaging include supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning, each serving distinct purposes (Hiran et al., 2021). Supervised learning involves training models on labeled datasets, enabling the AI system to associate patterns in medical images with their corresponding labels or diagnoses, making it valuable for tasks like image classification, object detection, and disease diagnosis (Verma et al., 2021). Unsupervised learning, on the other hand, focuses on finding patterns and structures in unlabeled data, which can lead to the discovery of hidden relationships or groupings within large datasets, potentially

unveiling new insights or discoveries in medical diagnostics and research (Raza & Singh, 2021). Semi-supervised learning algorithms combine elements of both supervised and unsupervised learning, leveraging the labeled data for guidance while also extracting information from the unlabeled data to improve the learning process, which is particularly beneficial when obtaining fully labeled data is costly or time-consuming (Chebli et al., 2018). While less commonly applied in medical imaging, reinforcement learning has the potential to play a significant role in dynamic and interactive systems, where an AI agent learns to make sequential decisions by interacting with an environment and receiving feedback in the form of rewards or penalties (Zhou et al., 2021).

In the realm of medical image analysis, the process of feature extraction plays a crucial role in detecting and classifying medical conditions. It involves identifying and selecting pertinent information from the input data to enhance the learning process and improve the performance of the model. Traditional machine learning approaches typically involve manually engineering features from the raw data, a process that often necessitates domain expertise and collaboration with clinicians to identify relevant visual patterns and characteristics that indicate specific medical conditions (Bektaş et al., 2018). Various techniques are employed, such as selecting important variables, transforming the data into a more suitable representation, or generating new features based on domain knowledge. Common feature types include texture features (capturing spatial arrangement of pixel intensities and texture patterns), shape features (obtained through contour analysis and morphological operations), and intensity features (statistical properties of pixel intensities) (Meenalochini & Ramkumar, 2021). However, not all extracted features carry equal importance for classification purposes. To address this, feature selection techniques like correlation-based feature selection (CFS), recursive feature elimination (RFE), and principal component analysis (PCA) are employed to identify the most relevant and discriminative features and reduce the dimensionality of the feature space, enhancing the efficiency and accuracy of the classification process (Jafari & Karami, 2023). Once the relevant features are selected, various machine learning classifiers are trained to classify medical images into normal or abnormal categories. These classifiers encompass a range of algorithms such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN), each bringing its unique strengths and characteristics to the task at hand (El_Rahman, 2021; Gayathri et al., 2013).

Support Vector Machines (SVMs) are powerful supervised machine learning algorithms used for classification and regression tasks. They find the optimal hyperplane that separates different classes of data points with the maximum possible margin. For linearly separable data, SVMs maximize the distance between the closest data points (support vectors) from each class. For non-linear data, the kernel trick maps data to a higher-dimensional space where it becomes linearly separable using kernel functions like linear, polynomial, RBF, or sigmoid. SVMs can handle high-dimensional data and are robust to overfitting by maximizing the margin between classes. They can handle binary and multi-class classification problems using strategies like one-vs-one or one-vs-rest. Limitations include computational expense for large datasets or non-linear kernels, and performance dependence on kernel function and parameter tuning (Azar & El-Said, 2014).

Several studies have applied SVM classification for breast cancer detection in mammograms using diverse methods. Chu et al. (2015) developed a CAD system employing techniques like morphological enhancement, SLIC segmentation, and under-sampled SVM ensembles, achieving high sensitivity (up to 98.55%) and low false positives (0.3-0.84 per image). De Nazare Silva et al. (2015) used an approach involving contrast enhancement, wavelet transform, shape descriptors, and SVM classification, reporting 92.31% sensitivity, 82.2% specificity, and 83.53% accuracy. De Sampaio et al. (2015) proposed a multi-stage method with breast density classification, micro-genetic mass segmentation, DBSCAN and texture analysis using phylogenetic trees, LBP, and SVM, achieving sensitivities of 83.7% (dense) and 92.99% (non-dense) with low false positives. Further SVM research includes Da Rocha et al. (2016) using LBP and ecology-inspired texture descriptors, Berbar (2018) exploring GLCM texture features and hybrid wavelet-contourlet methods, achieving up to 98.7% accuracy, and De Brito Silva (2020) categorizing masses based on geometric and topological characteristics using distance and surface maps, with up to 96.29% sensitivity and 91.05% specificity.

Decision trees (DTs) are popular machine learning algorithms used for classification and regression tasks. They are tree-like models that make decisions based on a hierarchical series of rules represented as branches. The fundamental idea is to recursively partition the input data based on feature values, creating a hierarchical structure of nodes (features) and branches (decision rules based on feature values), with leaf nodes representing the final output. Constructing a decision tree involves selecting the most discriminative features and determining optimal split points for each

feature. A key advantage is their interpretability - unlike "black box" algorithms, decision trees provide a clear and intuitive representation of the decision-making process, making them easy to understand, explain, and validate (Ghiasi & Zendehboudi, 2021). This is particularly important in domains like finance, healthcare, and law where interpretability is crucial. Decision trees can handle both numerical and categorical data, as well as missing data and outliers. As non-parametric models, they do not make assumptions about the underlying data distribution, making them flexible and applicable to a wide range of problems (Sathiyanarayanan et al., 2019). However, they are prone to overfitting when the tree becomes too complex and captures noise in the training data. To mitigate overfitting, techniques like pruning (removing branches that do not significantly contribute to accuracy) and ensemble methods (combining multiple decision trees to improve performance and reduce overfitting, e.g., random forests, gradient boosting) are employed. Decision trees are widely used in various applications, including credit risk analysis, fraud detection, spam filtering, image recognition, and medical diagnosis. They can be used as standalone models or as building blocks for more complex ensemble models. With their interpretability, flexibility, and ease of implementation, decision trees remain a popular choice among machine learning practitioners, especially when coupled with ensemble techniques to enhance their predictive power and robustness (Venkatesan & Velmurugan, 2015).

Random Forests (RF) are an ensemble learning algorithm that combines multiple decision trees to create a powerful and robust predictive model for classification and regression tasks. They are known for high accuracy and ability to handle high-dimensional data and missing values. The core idea is to construct a large number of decision trees, each trained on a different random subset of the training data (bagging/bootstrap aggregating) and different random subset of features (Nguyen et al., 2013). This introduces randomness that reduces correlation between individual trees, mitigating overfitting and improving generalization. During prediction, each tree makes a prediction, and the final output is determined by aggregating all predictions - majority voting for classification, and averaging for regression. Key advantages include handling high-dimensional data and complex feature interactions by considering only a random subset of features at each split. They are also relatively robust to outliers and noise, and provide estimates of feature importance useful for feature selection. However, Random Forests can be computationally expensive for large datasets and many trees, and lack the interpretability of individual decision trees since the prediction is an aggregate. Despite this, with increasing computational resources and need for

accurate robust models, Random Forests have become popular across various domains like finance, biology, computer vision, and natural language processing (Statnikov et al., 2008; Wang et al., 2020). In 2018, Dhahbi et al. addressed false positive masses detected by CAD systems in mammograms using a segmentation-free framework with Hilbert image representation, Kolmogorov-Smirnov distance, and maximum subregion descriptors. Multiple classifiers including Random Forest were employed on a large dataset of 10,168 regions from the DDSM, with Random Forest achieving the highest accuracy of 81.09% in distinguishing normal tissues from masses, though not statistically significantly better than SVM and Decision Trees.

Gradient boosting (GB) is a powerful machine learning ensemble technique that combines multiple weak learners, typically decision trees, to create a strong predictive model. It works by iteratively constructing an ensemble, where each new weak learner is trained to improve upon the errors of the previous ones by minimizing a loss function. This is done by computing the negative gradients of the loss and using them as targets for the next weak learner. Each subsequent learner predicts the residual errors from the previous ensemble, with its predictions added to the overall model with a small weight. The final model is a weighted sum of all weak learners (Pinheiro & Becker, 2024). Key advantages of GB include the ability to capture complex non-linear relationships, and resistance to overfitting due to the sequential additive training process acting as regularization. However, GB models can be sensitive to hyperparameters like learning rate, number of iterations, and tree depth - improper tuning can lead to over/underfitting. Regularization techniques like tree pruning and subsampling are used to mitigate this. GB has been successfully applied across various domains like computer vision, NLP, bioinformatics, and finance. Popular efficient implementations include XGBoost, LightGBM, and CatBoost, which incorporate additional features and optimizations (Kumar et al., 2022; Tabrizchi et al., 2020). In 2018, Eltoukhy et al. introduced an accurate breast CAD system using exact Gaussian-Hermite moment features fed into K-NN, random forests, and AdaBoost classifiers. Evaluated on IRMA and MIAS datasets via 10-fold cross-validation, it achieved impressive accuracy of 93.3% (IRMA) and 90.6% (MIAS), with AUC of 0.96 (IRMA) and 0.89 (MIAS), outperforming conventional methods in distinguishing normal vs abnormal lesions.

Logistic Regression (LR) is a popular statistical model used for binary classification tasks, where the goal is to predict the probability of an instance belonging to one of two classes (e.g. positive

or negative) based on its features or independent variables. The LR model estimates the probability of an instance belonging to the positive class using the logistic sigmoid function, which takes a linear combination of the input features weighted by their coefficients (learned during training) and maps it to a value between 0 and 1 representing the predicted probability (Ayer et al., 2010). During training, LR aims to find the optimal set of weights that maximizes the likelihood of the observed data, typically using techniques like Maximum Likelihood Estimation or gradient descent methods on the log-likelihood function. A key advantage of LR is its interpretability - the weights assigned to each feature directly represent the change in log-odds of the outcome for a one-unit increase in that feature, holding others constant. This makes LR valuable in domains where understanding individual feature impacts is crucial, like healthcare, finance, and social sciences (Ayer et al., 2010). Despite its simplicity, LR is a powerful binary classification technique, often used as a baseline or component in ensemble models. However, it assumes a linear relationship between features and log-odds of the outcome, which may not always hold in complex real-world scenarios where more flexible models like decision trees or neural networks may be preferable. Nonetheless, LR remains a valuable machine learning tool, particularly when interpretability and simplicity are prioritized over highly complex non-linear modeling capabilities (ABD ALMALEKI et al., 2004).

Naive Bayes (NB) is a family of simple yet efficient probabilistic classifiers based on Bayes' theorem and the assumption that features are conditionally independent given the class label. Despite this "naive" independence assumption being rarely true in reality, NB classifiers often perform remarkably well, especially for text classification and high-dimensional data (Kamel et al., 2019). NB works by computing the posterior probability of each class given the feature values, using Bayes' theorem and the assumption that the likelihood of observing the features can be calculated as the product of the individual feature likelihoods due to independence. The class with the highest posterior probability is predicted. Key advantages of NB include simplicity, efficiency in handling high-dimensional and missing data, and often competitive performance. Different variations like Gaussian NB (continuous data), Multinomial NB (discrete counts), and Complement NB (imbalanced data) exist. While NB struggles with complex non-linear relationships where more sophisticated models may be preferred, it remains a valuable tool when interpretability, efficiency and ease of implementation are priorities (Karabatak, 2015).

In 2018, Chakraborty et al. presented an approach for automatically detecting and diagnosing mammographic masses as benign or malignant breast cancer indicators. Their iterative thresholding and radial region growing method aimed to detect masses based on tissue pattern orientation changes. Multi-resolution orientation analysis then categorized masses. Evaluating on the DDSM dataset with 450 benign, 440 malignant, and 410 normal images, the algorithm achieved 85% sensitivity with 1.4 false positives per image for mass detection. For diagnosis, it had 0.92 AUC and 83.3% accuracy in distinguishing benign vs malignant masses.

The K-Nearest Neighbors (KNN) algorithm is a simple yet powerful non-parametric method used for classification and regression tasks in machine learning. It is an instance-based learning algorithm that makes predictions based on the similarity of new instances to existing instances in the training data (Khorshid & Abdulazeez, 2021). For classification, KNN classifies a new data point by:

1. Calculating the distance to all points in the training set using a distance metric
2. Selecting the k nearest neighbors based on distance
3. Assigning the class that is most common among the k nearest neighbors (majority vote)

For regression, it computes the average of the target values of the k nearest neighbors as the predicted value (Alarabeyyat & Alhanahnah, 2016). Key advantages of KNN include simplicity, ability to handle non-linear decision boundaries without assumptions on data distribution, and effectiveness for complex classification/regression problems. Limitations include computational expense for large datasets, sensitivity to distance metric and k value selection, irrelevant features, and curse of dimensionality issues. Despite limitations, KNN remains popular across domains like pattern recognition, image classification, and recommender systems, often used as a baseline or in ensembles. Its simplicity and interpretability aid exploratory data analysis (Khorshid & Abdulazeez, 2021).

In 2015, Dhahbi et al. addressed feature extraction challenges in breast cancer CAD systems. Previous curvelet transform methods performed well but resulted in high-dimensional features. They proposed a novel method based on discrete curvelet transform and moment theory, computing first-order moments from curvelet coefficients, with t-test ranking for feature selection. Using a KNN classifier on mini-MIAS and DDSM datasets, their method achieved 91.27% accuracy for abnormality detection and 81.35% for malignancy on mini-MIAS with compact

feature sets, statistically outperforming other curvelet-based methods on DDSM while providing dimensionality reduction.

Artificial Neural Networks (ANNs) are powerful machine learning models inspired by biological neural networks. They consist of interconnected artificial neurons organized into layers that can learn complex patterns by processing input data and adjusting connection weights during training (Azar & El-Said, 2013). Key components include weights determining connection strengths, activation functions introducing non-linearity, input/output layers, and hidden layers performing intermediate computations. Deeper networks with more hidden layers are generally more powerful but prone to overfitting. ANNs use optimization algorithms like gradient descent to iteratively adjust weights based on prediction errors during training. They excel at learning complex non-linear relationships and can handle supervised and unsupervised tasks across domains like computer vision, NLP, speech, and games. However, training deep ANNs requires substantial computational resources, data, and careful regularization (Abbass, 2002; Thein & Tun, 2015).

For breast cancer detection, studies employed techniques like Particle Swarm Optimized Wavelet Neural Networks (2014) using Laws texture features, and enhanced CAD models (2019) combining discrete transforms, PCA/LDA dimensionality reduction, and optimized extreme learning machines. Evaluations showed high accuracies around 96-100% on mammogram datasets like MIAS and DDSM. In breast cancer classification, Lim and Er (2004) used generalized dynamic fuzzy neural networks with texture parameters, achieving 0.868 AUC and high true positive rate. Xie et al. (2016) proposed an ELM-based CAD system with image segmentation and SVM-ELM feature selection, reporting 96.02% average accuracy in distinguishing malignant vs benign masses.

An experiment using the logistic regression (LR) classifier on the Wisconsin Breast Cancer Diagnosis (WBCD) dataset showed that proper feature selection, such as using maximum perimeter and maximum texture, can improve LR's accuracy up to 96.5%. However, another study found that decision trees performed slightly better than LR, although both achieved high accuracy rates. Neural network classifiers like multi-layer perceptron (MLP) and radial basis function networks (RBF) have been widely studied. MLP outperformed RBF and is considered one of the most effective classifiers, despite lacking interpretability compared to methods like Naive Bayes (NB). NB, built on Bayes' theorem, demonstrated good performance and interpretability despite

violating its independence assumption. Techniques like weighted NB aimed to overcome regular NB's drawbacks and achieved better accuracy. Support vector machines (SVM), particularly linear SVM, topped ANN and NB in accuracy on the WBCD dataset and showed precise diagnosis capacity. When using features extracted from mammogram images with image processing, backpropagation neural networks (BPNN) exceeded 93% accuracy with no more than 240 features, outperforming LR (Alshayeji et al., 2022).

The ensemble random forest (RF) classifier, combining multiple decision trees, enhanced accuracy. RF outperformed NB and k-nearest neighbors (KNN) on the WBCD data, although KNN achieved higher overall accuracy, precision, and F1 score. ANN with feature selection methods showed considerable performance improvement over SVM, NB, and decision trees. Comparative studies on the WBCD dataset found that KNN achieved the highest accuracy among nine models, including LR, Gaussian NB, SVM, decision trees, RF, XGBoost, and gradient boosting, for supervised learning. LR performed best for semi-supervised learning. On a smaller dataset, RF outperformed XGBoost, though larger datasets are needed for reliable comparisons. Ensemble learning approaches like stacking, boosting, and bagging, which combine individual classifiers, have been shown to improve performance compared to single models. Common ensembles use ANN, SVM, and tree-based classifiers, often evaluated on the WBCD dataset, which is widely trusted in the research community for breast cancer classification (Alshayeji et al., 2022).

2.2 Deep Learning Applications in Breast Cancer Detection and Classification

Deep learning (DL) is an advanced branch of machine learning that utilizes artificial neural networks (ANNs) with multiple hidden layers. This architecture allows DL algorithms to learn intricate feature representations directly from raw data, leading to improved performance across various tasks (Ongsulee, 2017). Popular DL architectures include deep belief networks (DBNs), recurrent neural networks (RNNs), autoencoders, and convolutional neural networks (CNNs), and each suited for different data types and applications. CNNs have gained significant popularity in image-based tasks, demonstrating performance that surpasses human experts in specific domains (Sarker, 2021). A typical CNN architecture consists of an input layer, multiple hidden layers, and an output layer. The input layer receives raw image data with dimensions representing height, width, and depth (channels or modalities). The hidden layers perform specific computations to

extract meaningful features, while the output layer produces final predictions or classifications (Sun et al., 2019).

In the initial layers, CNNs learn low-level features such as edges, textures, and patterns from the input image through convolution operations. Convolutional layers convolve the input with a set of filters (kernels), sliding across the image and computing dot products between filter weights and corresponding input regions. The resulting feature maps capture different aspects of the input data. Nonlinear activation functions, like the rectified linear unit (ReLU), introduce nonlinearity into the network by replacing negative pixel values with zero. This operation introduces sparsity and enables the network to learn more complex and abstract features from the input data (Alzubaidi et al., 2021).

Pooling layers play a crucial role in CNNs by enhancing computational efficiency and reducing spatial dimensions. Techniques like max pooling and average pooling help control the number of parameters, prevent overfitting, and improve computational speed. Max pooling retains the maximum pixel value from a region, capturing the most prominent features, while average pooling retains the average pixel value, providing a more generalized representation (Jie & Wanda, 2020; Sun et al., 2017). In the fully connected (FC) layer, neurons establish dense interconnections with all activations from the preceding layer. This interconnectedness allows the network to learn complex patterns and relationships within the data. The FC layer generates a feature vector representing the high-level representations learned by the network, which serves as input for the final classification step. The output layer of a CNN applies the softmax activation function to perform classification. The softmax function transforms the non-normalized output into a set of probability values, with each value representing the likelihood of the input belonging to a specific class. The input is then classified into the category with the highest probability (Basha et al., 2020).

Deep learning techniques, particularly convolutional neural networks (CNNs), have revolutionized breast cancer detection and classification from mammogram images. Unlike traditional methods relying on handcrafted features, CNNs can automatically learn hierarchical features directly from the images through their convolutional layers. This ability eliminates manual feature engineering and enables the model to acquire complex representations crucial for accurate classification. In the deep learning pipeline, the convolutional layers extract features by applying filters that capture local patterns at various scales. The learned features are then mapped to class probabilities by fully

connected layers using activation functions like softmax or sigmoid (Arora et al., 2020). Model training involves optimizing network parameters on labeled data through backpropagation and gradient descent to minimize classification error. Architectures like AlexNet, VGGNet, GoogLeNet, and ResNet have exhibited exceptional performance in this domain. Evaluation metrics such as accuracy, sensitivity, specificity, precision, recall, F1-score, ROC curves, and AUC are employed to assess the trained model's performance across different thresholds. Numerous studies have leveraged deep learning for breast cancer detection from mammograms, demonstrating its potential to enhance early detection rates and assist radiologists (Warnecke et al., 2020).

In recent years, researchers have made significant strides in developing computer-aided diagnosis (CAD) systems for breast cancer detection and classification using deep learning techniques on mammogram images. A notable study by Dhungel et al. (2017) presented an integrated three-stage approach involving mass detection, segmentation, and classification. Their methodology employed m-DBN, GMM, CNN, and RF for candidate generation and false positive reduction during mass detection. Segmentation utilized CRF and the Chan-Vese active contour method, while classification involved pre-training and fine-tuning a CNN model. When tested on the INbreast dataset, this approach achieved a 90% mass detection rate, 0.85 segmentation accuracy (Dice index), 0.98 sensitivity, and 0.7 specificity for classification. Object detection models like YOLO have also been explored for simultaneous mass detection and classification. Al-masni et al. (2018) introduced a YOLO-based CAD system that demonstrated remarkable performance, with 99.7% accuracy in detecting mass locations and 97% accuracy in distinguishing between benign and malignant lesions on the DDSM and augmented dataset. Aly et al. (2020) compared different YOLO architectures, finding that YOLO-V3 with k-means clustering achieved 89.4% detection rate and up to 95.5% classification accuracy when combined with ResNet and InceptionV3 networks.

Other studies have utilized two-stage object detectors like Faster R-CNN and RetinaNet. Ribli et al. (2018) proposed a Faster R-CNN CAD system that achieved an impressive AUC of 0.95 on the INbreast dataset and secured second place in the DREAM challenge with an AUC of 0.85. Jung et al. (2018) developed a mass detection model based on RetinaNet, consistently outperforming more complex models and achieving 98% sensitivity with only 1.3 false positives per image on the

INbreast and GURO datasets. Additionally, researchers have explored the potential of convolutional neural networks (CNNs) for mass classification. Shen et al. (2019) developed an end-to-end CNN that effectively utilized training datasets with or without lesion annotations, achieving AUCs of 0.88-0.98 on the CBIS-DDSM and INbreast datasets. Gnanasekaran et al. (2020) presented a CNN model for classifying mammograms into benign, malignant, and normal categories, surpassing pre-trained networks like AlexNet and VGG16 with up to 98.32% accuracy and an AUC of 0.98 on combined datasets.

The increasing global demand for early breast cancer detection has driven extensive research into improving the accuracy of computer-aided diagnosis (CAD) systems. These systems have become crucial tools for detecting and distinguishing various breast abnormalities. In 2016, Abdel-Zaher and Eldeib developed a CAD scheme using deep belief networks and backpropagation. Tested on the Wisconsin Breast Cancer Dataset, it achieved a remarkable accuracy of 99.68%, surpassing previous studies. The proposed system provides an effective breast cancer classification model, demonstrating an accuracy of 99.7% on 683 digitized samples using data augmentation techniques.

Wang et al. (2018) tackled the challenge of false positives in detecting clustered microcalcifications (MCs) by developing a context-sensitive deep neural network (DNN). The DNN considers both local MC features and surrounding tissue background. Evaluated on 292 mammograms, it showed significantly higher accuracy in detecting individual MCs and clusters compared to MC-based detectors, achieving 85.2% accuracy on 736 mammograms from the BCDR-F03 dataset. Arora et al. (2020) achieved 88% accuracy using deep ensemble transfer learning and a neural network classifier for automatic feature extraction and classification of mammogram images. Their computer-assisted mammography approach demonstrated the robustness of the proposed CADx system for breast cancer classification. Al-Antari et al. (2020) proposed an integrated CAD system utilizing deep learning techniques for breast lesion detection and classification. The YOLO detector achieved high detection accuracies of up to 99.17% and an F1-score of 99.28% on the DDSM and INbreast datasets. The deep learning classifiers (CNN, ResNet-50, and InceptionResNet-V2) also demonstrated promising average overall accuracies, with InceptionResNet-V2 reaching 97.5% and 95.32% on the DDSM and INbreast datasets, respectively.

In summary, deep learning techniques, such as convolutional neural networks (CNNs), harness the capabilities of multiple hidden layers to extract intricate features from raw data. The architecture of a CNN, comprising convolutional layers, pooling layers, and fully connected layers, enables the network to learn hierarchical representations and excel in tasks like image recognition, object detection, and image classification. Activation functions and pooling techniques further enhance the performance and efficiency of the network. However, deep learning has its limitations (Abdelrahman et al., 2021). Firstly, it requires a substantial amount of labeled training data, which can be challenging and costly to acquire, particularly in specialized domains. Secondly, the computational demands of deep learning models often necessitate high-end resources like Graphical Processing Units (GPUs), which may not be accessible to all researchers or organizations. Lastly, deep learning models can struggle to generalize well to real-time datasets, leading to diminished performance when applied to new and unseen data (Becker et al., 2017).

To overcome these limitations, transfer learning (TL) has emerged as a new paradigm in deep learning (Wahab et al., 2019). TL involves transferring knowledge learned from one task to a related task, leveraging previous knowledge to enhance performance on the new task. The steps involved in TL for medical imaging tasks are as follows: 1) Select a pre-trained model from established architectures that have been trained on a large dataset. 2) Customize the last layers of the pre-trained model to suit the specifics of the new task's data. 3) Retrain the model using the new medical dataset to adapt it to the specific characteristics of the data. 4) Evaluate the performance of the retrained network using test data and appropriate metrics to ensure its suitability for the intended medical imaging task. Pretrained networks are also utilized for deep feature extraction, with earlier layers providing low-level information and deeper layers providing high-level information. Optimal hyperparameters and performance evaluation metrics derived from the confusion matrix are crucial for building high-performance deep learning systems (Gupta & Chawla, 2020). For breast cancer detection and classification, various pre-trained CNN models, including AlexNet, VGGNet, GoogLeNet, ResNet, and DenseNet, have demonstrated excellent performance in image detection and classification tasks. These architectures differ in terms of layer numbers, filter sizes, and activation functions, providing flexibility in designing models based on specific problems and datasets (Gonçalves et al., 2021).

AlexNet, introduced in 2012 by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, is a groundbreaking convolutional neural network architecture that revolutionized deep learning and computer vision. It achieved remarkable performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by outperforming traditional computer vision techniques. The network comprises eight layers, including five convolutional layers and three fully connected layers. The input is a 227x227-pixel image. The first layer applies 96 filters of size 11x11 with a stride of 4 pixels, followed by a ReLU activation function and a max-pooling layer with a 3x3 filter and a stride of 2 pixels. The second layer applies 256 filters of size 5x5 with a stride of 1 pixel, followed by ReLU activation and max-pooling with the same parameters. The third, fourth, and fifth layers apply 384, 384, and 256 filters of size 3x3 with a stride of 1 pixel, respectively. ReLU activation functions are applied after these layers, with no pooling. After the convolutional layers, AlexNet includes three fully connected layers with 4096 neurons each, and the final layer has 1000 neurons representing the object categories in the ImageNet dataset. A softmax layer produces the output probabilities for each class (Alom et al., 2018). AlexNet introduced several innovations, such as ReLU activation functions to address the vanishing gradient problem, dropout regularization to reduce overfitting, and data augmentation techniques to expand the training dataset. Its success demonstrated the superiority of deep convolutional neural networks over traditional methods and paved the way for further advancements in deep learning and its application in computer vision tasks like object detection, semantic segmentation, and image generation (Cao et al., 2020).

The VGGNet, introduced by researchers at the University of Oxford in 2014, is a popular deep convolutional neural network architecture known for its simplicity and impressive performance on the ImageNet dataset. It follows a consistent design principle, consisting of convolutional layers, max-pooling layers, and fully connected layers. The convolutional layers use 3x3 filters with a stride of 1, while the max-pooling layers have 2x2 filters with a stride of 2. This design choice allows for more non-linearities and reduces the number of parameters. VGGNet offers two variants: VGG-16 and VGG-19, referring to the number of weight layers. VGG-16 includes 13 convolutional layers organized into five blocks, with increasing numbers of filters in each block. VGG-19 is similar to VGG-16 but has 16 convolutional layers organized into five blocks. Both variants end with three fully connected layers (Muhammad et al., 2018). The simplicity and uniformity of VGGNet make it easy to understand but result in a larger number of parameters and

increased computational complexity compared to other architectures. Nevertheless, VGGNet achieved impressive performance on the ImageNet dataset and highlighted the effectiveness of deeper neural networks. It has been widely used as a backbone for various computer vision tasks and has inspired further research into deeper and more efficient architectures (Zhang et al., 2020).

GoogLeNet, introduced by Google researchers in 2014, is a groundbreaking convolutional neural network architecture known as Inception. Its goal was to address computational complexity and overfitting issues while achieving top performance in computer vision tasks. The Inception module is a key component of GoogLeNet, combining convolutional layers with different filter sizes and a max-pooling layer. This module allows the network to capture features at multiple scales simultaneously, enabling it to learn richer representations. The use of 1x1 convolutions helps reduce computational cost. GoogLeNet consists of stacked Inception modules with interspersed max-pooling layers for downsampling. The network begins with a traditional convolutional layer, followed by max-pooling and two Inception modules. It then repeats this pattern, increasing the number of filters to capture higher-level features (Anand et al., 2020). An innovation of GoogLeNet is the inclusion of auxiliary classifiers, small fully connected layers attached to intermediate Inception modules, which aid in training and prevent overfitting. Another notable feature is the use of global average pooling instead of fully connected layers, reducing parameters and overfitting. GoogLeNet achieved impressive performance in the ImageNet challenge in 2014, surpassing other models while being computationally efficient. Its success highlighted the effectiveness of multi-scale feature extraction, dimensionality reduction, and auxiliary classifiers in deep neural networks. GoogLeNet has inspired architectural variations and has been widely adopted in computer vision applications, driving further advancements in deep learning architectures (He, 2020).

ResNet, introduced by Microsoft Research in 2015, is a groundbreaking deep convolutional neural network architecture designed to address the degradation problem in deep networks. It introduces residual connections, allowing the network to bypass layers and perform identity mappings. The core idea is to learn residual functions with respect to the layer inputs. The building block of ResNet is the residual block, consisting of convolutional layers, batch normalization, and ReLU activation. The output is added to the input through a skip connection, creating a residual mapping. ResNet can have hundreds or thousands of layers thanks to these connections. Different ResNet

architectures were introduced, ranging from 34 to 152 layers (Targ et al., 2016). The network uses different sizes of convolutions for feature extraction and employs batch normalization for stabilization. Stride convolutions or pooling layers downsample feature maps, and the residual connections enable the training of deeper networks. ResNet achieved improved performance on various computer vision tasks and has been widely adopted and extended in the field of deep learning and computer vision (Gao et al., 2022).

DenseNet, introduced in 2017, is a deep convolutional neural network architecture known for its dense connections. These connections enable feature maps from previous layers to be directly connected to subsequent layers, promoting better feature propagation and reuse. DenseNet consists of multiple dense blocks, each with convolutional layers, batch normalization, and ReLU activation. The feature maps from preceding layers are concatenated and used as input to the next layer within each dense block. This dense connectivity facilitates information propagation and alleviates the vanishing gradient problem. Transition layers between dense blocks downsample feature maps using convolution and pooling operations (Zhu & Newsam, 2017). DenseNet achieves parameter efficiency by reusing feature maps, leading to state-of-the-art performance with fewer parameters compared to traditional networks. Different variants, such as DenseNet-121 and DenseNet-264, vary in depth. DenseNet has demonstrated superior results on benchmark datasets for image classification, object detection, and semantic segmentation. Its dense connectivity has improved information flow and gradients, contributing to advancements in computer vision and deep learning (Shaik & Kirthiga, 2021).

Ting et al. (2019) implemented a deep CNN for BC-lesion classification. This network consisted of 1 input layer, 28 hidden layer, and 1 output layer. Overfitting was avoided using the feature-wise-data augmentation (FWDA) algorithm. Their proposed method sequentially achieved 89.47%, 90.50%, and 90.71% for sensitivity, accuracy, and specificity, respectively. Toğçar et al. (2020) proposed the BreastNet, which consisted of convolutional, pooling, residual, and dense blocks, and it was capable of extracting the most effective features from breast images. BreastNet achieved better results than AlexNet, VGG-16, and VGG-19 models as its accuracy approached 98.80%. Abbas (2016) presented a multi-layer DL architecture for classifying benign and malignant regions in breast images. This network consisted of four phases for extracting invariant features, which were transformed into deep-invariant features, and learning features for making

the final decision. In 2016, the MIAS dataset was used and achieved a 92%, 84.2%, 91.5%, and 0.91 for sensitivity, specificity, accuracy, and AUC, respectively. Using the same dataset, Sha et al. (2020) presented a method for automatic detection and classification of the cancerous region in breast images. Their proposed method was based on CNNs and the grasshopper optimization algorithm. The results showed that this proposed method was capable of achieving 96%, 93%, and 92% for sensitivity, specificity, and accuracy, respectively.

Charan et al. (2018) trained a CNN for BC detection. Their proposed CNN consisted of six convolution layers, four average-pooling layers, and three fully-connected layers (FCLs). They used a size of 224 × 224 for the input image and the Softmax (SM) function to apply the classification results. The overall accuracy of this network was 65%, which was obtained using the MIAS database. In (2019), Wahab et al. exploited a pre-trained CNN and transferred its learned parameters to another CNN for mitoses classification. Their proposed method achieved 0.50, 0.80, and 0.621 for precision, recall, and F-measure, respectively. In addition, for multi-class BC-classification purposes, Lotter et al. (2019) proposed a model in which the features were extracted using a pre-trained ResNet50 network. Their model was capable of classifying lesions into five classes: mass, calcifications, focal asymmetry, architectural distortion, or no lesion. Their model achieved 96.2, 90.9, and 0.94 for sensitivity, specificity, and AUC, respectively. Jiang et al. (2017) achieved better BC-classification accuracy in the case of TL from a pre-trained network in building networks from scratch. The accuracy approached 0.88 using GoogleNet and 0.83 using AlexNet on the film mammography number 3 (BCDR-F03) dataset. Khan et al. (2019) implemented a model in which the breast-image features were extracted using pre-trained CNN architectures, namely, GoogleNet, VGGNet, and ResNet. The model's accuracy, which approached 97.525%, was evaluated using a standard benchmark dataset. Cao et al. (2018) improved the performance of TL for BC-classification without any fine-tuning on the source network layers (ResNet-125). Instead, they used random forest dissimilarity for combining various feature groups. The ‘‘ICIAR 2018’’ dataset was used, and the classification accuracy was improved to 82.90%. Deniz et al. (2018) fine-tuned the last three layers in the AlexNet and VGG16 models to classify breast tumors on the BreaKHis dataset. Their model achieved better accuracy than five other methods as it approached 91.37%. In the same dataset, Celik et al. (2020) pre-trained the DenseNet161 model and achieved 92.38% and 91.57% for the F-score and accuracy, respectively.

From the provided information, it is clear that machine learning remains relevant and widely utilized in the field of breast cancer detection and classification. Although deep learning has gained considerable attention and popularity in recent times, it does not imply that machine learning techniques are no longer employed. The emphasis on deep learning in the mentioned studies may stem from its ability to automatically extract features from raw data, such as mammogram images, eliminating the need for manual feature engineering. However, traditional machine learning algorithms, including support vector machines and random forests, continue to be utilized in various research studies and practical applications. It is crucial to recognize that the choice between machine learning and deep learning relies on the specific problem, available data, and desired outcomes. Each approach possesses its own strengths and limitations, and researchers often select the most appropriate method based on the requirements of their study.

Table 4: A compilation of prior research studies discussing the utilization of artificial intelligence for the detection and classification of breast cancer from mammogram images, along with the performance metrics attained in each study.

Reference	**Database**	**Type of images**	**Dataset**	**Classifier**	**Results**			
					ACC [%]	**SEN/SPEC [%]**	**AUC**	**Other**
Machine Learning								
Tai, Chen and Tsai (2014)	Open Access (DDSM)	Digitized	358 images	LDA	–	90.3/-	0.98	–
Dheeba et al. 2014 (2014)	Restricted	Digitized	216 images	Particle Swarm Optimized Wavelet Neural Network (PSOWNN)	93.7	94.2/92.1	0.97	–
Chu et al. (2014)	Open Access (DDSM)	Digitized	474 images	SVM	81.4	93.4/78.2	–	0.84 FPi

de Nazare Silva et al. (2015)	Open Access (DDSM)	Digitized	599 images	SVM	83.5	92.3/82.2	0.8	1.1 Fpi
Dhahbi, Barhoumi and Zagrouba (2015)	Open Access (mini-MIAS)	Digitized	252 images	k-NN	91.3	–	0.99	–
De Sampaio et al. (2015)	Open Access (DDSM)	Digitized	1727 images	SVM	–	93/- 83.7/-	–	0.15 FPi 0.19 FPi
Chakraborty, Midya and Rabidas (2018)	Open Access (DDSM)	Digitized	1300 images	NB	–	85/-	–	1.4 FPi
Eltoukhy et al. (2018)	Open Access (IRMA MIAS)	Digitized	1516 ROIs 256 ROIs	AdaBoost	93.3 90.6	–	0.96 0.89	–
Dhahbi et al. (2018)	Open Access (DDSM)	Digitized	10168 ROIs	RF	81.1	–	–	–
Mohanty et al. (2019)	Open Access (MIAS DDSM)	Digitized	314 images 1500 images	Improved Grey Wolf Optimization-based ELM (IGWO-ELM)	100 99.5	–	1 0.99	–
Lim and Er (2004)	Open Access (DDSM)	Digitized	343 images	Generalized Dynamic Fuzzy NN (GDFNN)	70	–	0.87	–

Tahmasbi et al. (2011)	Open Access (MIAS)	Digitized	322 images	MLP	96.4	–	0.98	–
Xie et al. (2016)	Open Access (MIAS DDSM)	Digitized	322 images 2620 patients	NN (Extreme Learning Machine)	96 95.7	96.3/94.3 94.9/97.2	0.97 0.97	–
Choi et al. (2016)	Open Access (DDSM)	Digitized	2743 ROIs 514 ROIs	Ensemble	–	–	0.93	–
da Rocha et al. (2016)	Open Access (DDSM)	Digitized	1155 ROIs	SVM	88.3	85/91.9	0.88	–
Danala et al. (2018)	Restricted	Digital	111 patients	MLP	78.4	80.8/72.7	0.85	–
Seryasat and Haddadnia (2018)	Open Access (mini-MIAS DDSM)	Digitized	55 images 240 images	Ensemble	94.8 92	–	0.96 0.94	–
Berbar (2018)	Open Access (DDSM MIAS)	Digitized	1024 images 291 images	SVM	98.7 97.9	99.2/- 96.1/-	0.98 0.88	–
de Brito Silva (2020)	Open Access (DDSM)	Digitized	794 ROIs	SVM	90.2	91/89.9	0.96	–
Deep Learning								

Dhungel et al. (2015)	Open Access (DDSM-BCDR INbreast)	Both	115 cases 79 cases	CNN (cascade of DL and RF)	–	96/- 75/-	–	1.2 FPi 4.8 Fpi
Dhungel et al. (2017)	Open Access (INbreast)	Digital	410 images	CNN	90	–	–	1 FPi
Charan et al. (2018)	Open Access (MIAS)	Digitized	322 images	CNN	65	–	–	–
Ribli et al. (2018)	Both (DDSM INbreast)	Both	2620 images 847 images 410 images	R-CNN	90	–	0.95	0.3 FPi
Agrawal et al. (2018)	Open Access (MIAS)	Digitized	322 images	Voting	80	–	–	–
Jung et al. (2018)	Open Access (INbreast)	Digital	410 images 222 images	CNN (RetinaNet)	–	98/-	–	1.3 Fpi
Gao et al. (2018)	Both (INbreast)	Digital	49 patients 89 patients	Shallow-Deep CNN (SD-CNN)	90	83/94	0.92	–

Al-masni et al. (2018)	Open Access (DDSM)	Digitized	600 images	CNN (YOLO)	99.7	83/94	0.97	–
Shen et al. (2019)	Open Access (CBIS-DDSM INbreast)	Both	2478 images 410 images	CNN	–	86.1/80.1 86.7/96.1	0.91 0.98	–
Zeiser et al. (2020)	Open Access (DDSM)	Digitized	7989 images	CNN (U-NET)	85.9	92.3/80.5	0.86	–
Aly et al. (2020)	Both (INbreast)	Digital	410 images	CNN (YOLO-V3)	89.4	–	–	–
Arevalo et al. (2015)	Open Access (BCDR-F03)	Digitized	736 images	CNN	–	–	0.86	–
Jiao et al. (2016)	Open Access (DDSM)	Digitized	600 images	SVM	96.7	–	–	–
Abdel-Zaher and Eldeib (2016)	Open Access (Wisconsin database)	Digitized	683 samples	CNN (Deep Belief Networks)	99.7	100/99.5	–	–
Al-antari et al. (2018)	Open Access (INbreast)	Digital	410 images	CNN	95.6	97/92.4	0.95	–
Wang et al. (2018)	Open Access	Digitized	736 images	Deep NN based on Multi-View data	85.2	–	0.89	–

	(BCDR-F03)							
Al-masni et al. (2018)	Open Access (DDSM)	Digitized	600 images	FC-NN	97	100/94	0.96	–
Arora et al. (2020)	Open Access (CBIS-DDSM)	Digitized	2620 images	DNN	88	91/-	0.88	–
Gnanasekaran et al. (2020)	Both (MIAS DDSM)	Digitized	322 images 1416 images 202 images	CNN	98.3	–	0.98	–
Al-antari et al. (2020)	Open Access (DDSM INbreast)	Both	600 images 410 images	CNN (InceptionResNetV2)	97.5 95.3	–	–	–
Barnett et al. (2021)	Restricted	Digitized	1136 images	DNN (Interpretable DL model)	83	–	–	–

2.3 Applications of Hybrid Artificial Intelligence Learning in Breast Cancer Detection and Classification

With the growing body of research and technological advancements, a novel approach known as hybrid learning has emerged, aiming to integrate machine learning and deep learning techniques to achieve more efficient and accurate diagnostic models. This approach has garnered significant interest and support from researchers and developers due to its demonstrated effectiveness in the field. Researchers are actively conducting studies on this approach, and among them are the following notable contributions. Several studies have explored using deep convolutional neural

networks (CNNs) and transfer learning for breast cancer classification from mammograms. Aarti Bokade and Ankit Shah (2021) proposed a method utilizing pre-trained CNNs like VGG16, VGG19, and ResNet-50 for feature extraction, which were then classified by a random forest model. This approach achieved high accuracies ranging from 80-99.6% across different datasets.

Dhungel et al. (2015) introduced a cascade framework combining deep belief networks, CNNs, and random forests for mass detection, reducing false positives while maintaining high sensitivity. Their method outperformed others on DDSM-BCRP and INbreast datasets. One study proposed representing global/local impressions of masses using high/mid-level CNN features along with the original images, combining predictions from multiple classifiers. Feature visualization showed deep features enhanced classification performance. Essam H. Houssein et al. (2022) developed IMPA-ResNet50, integrating ResNet50 and an optimization algorithm (IMPA), achieving 98.88% and 98.32% accuracies on MIAS and CBIS-DDSM datasets respectively. Dina A. Ragab (2021) fine-tuned pre-trained CNNs for binary classification, extracted features for SVM classifiers, and combined features from multiple CNNs like AlexNet and ResNet. Using PCA reduced features, this CAD system reached up to 97.9% accuracy. In 2019, Dina A. Ragab et al. utilized AlexNet features and an SVM to classify masses from segmented ROIs, achieving up to 88% accuracy on DDSM and 87.7% on CBIS-DDSM, outperforming prior methods. Hasan Nasir Khan et al. (2019) proposed a multi-view CNN using images from CC, MLO, LCC, and RCC views with early fusion, outperforming single-view CNNs on the CBIS-DDSM dataset. Runyu Song et al. (2020) combined deep CNN features with texture features like GLCM and HOG, using XGBoost to achieve 92.8% accuracy, higher than using individual feature types. Yu-Dong Zhang et al. (2021) developed BDR-CNN-GCN, combining a CNN and graph CNN, exhibiting superior malignant lesion detection compared to existing methods. Xiang Yu et al. (2019) fine-tuned DenseNet201 on mammograms using transfer learning, achieving 92.73% diagnostic accuracy in their semi-automated CAD system. Simon Hadush Nrea (2020) designed a CNN model with elements from Faster R-CNN for mass detection and classification, outperforming a VGG-based Faster R-CNN model on a dataset from Ethiopia.

Denan Lin (2020) proposed the DLA-EABA method combining a CNN for deep feature extraction with AdaBoost classification using handcrafted features, reaching 97.2% accuracy. L.S. (2019) developed a two-step algorithm training CNNs like ResNet-50 and VGG-16 on lesion patches,

then full mammograms, achieving over 0.85 AUC and performance comparable to clinicians. Vaira Suganthi Gnanasekaran (2020) introduced a CNN model tailored for breast mass classification, outperforming AlexNet and VGG16 with up to 98.32% accuracy after fine-tuning on a merged dataset. These studies demonstrate significant progress in developing accurate deep learning and machine learning models for breast cancer detection and classification from mammograms, with some matching or exceeding clinician performance.

This study aims to develop a hybrid AI approach that leverages the strengths of both deep learning and machine learning to enhance the accuracy and efficiency of breast cancer detection and classification from mammograms. The methodology involves several key steps. Firstly, a comprehensive dataset of mammogram images will be collected from reputable sources, ensuring diversity in terms of patient demographics, breast tissue types, and pathology. Preprocessing techniques will be applied to standardize the images and enhance their quality, thus preparing them for feature extraction and model training. Next, deep learning algorithms, such as convolutional neural networks (CNNs), will be employed for feature extraction and representation learning from the mammogram images. These learned features will then be fed into machine learning classifiers, such as support vector machines (SVM), random forests or any kind of machine learning, to further refine the classification process and improve the overall accuracy. Furthermore, the performance of the proposed hybrid AI model will be rigorously evaluated using metrics such as sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Comparative analyses with existing approaches and clinical experts' assessments will also be conducted to validate the effectiveness and clinical relevance of the developed model.

Chapter Three: Methodology

In this comprehensive retrospective analytical study, our objective was to examine the effectiveness of hybrid artificial intelligence models in the classification of mammogram images into benign and malignant cancers. To accomplish this, we meticulously gathered a diverse dataset of mammogram images. The first step of our methodology involved preprocessing the images and ensuring their suitability for analysis. We then proceeded to extract informative features from the mammogram images using two state-of-the-art deep learning models: VGG16 and DenseNet121. These models are renowned for their ability to capture intricate patterns and structures within medical images. After feature extraction, we employed a range of machine learning algorithms to perform the classification task. These algorithms included support vector machine (SVM), random forest, gradient boosting, and logistic regression. By leveraging the strengths of each algorithm, we aimed to enhance the overall predictive performance of our models.

The classification process was conducted in three stages, allowing for a comprehensive assessment of the models' capabilities. At each stage, the models were trained using a combination of labeled benign and malignant mammogram images. Following training, the models were evaluated using various performance metrics, including accuracy, precision, recall, f1-score, and the area under the receiver operating characteristic curve (AUC). These metrics provided valuable insights into the models' accuracy, ability to correctly identify positive cases, ability to avoid false positives, and overall discriminatory power. Additionally, we measured the computational efficiency of the models by recording the time taken for classification in seconds (TIME). This allowed us to evaluate not only the accuracy of the models but also their practical feasibility in real-world scenarios. By adopting this rigorous methodology, we aimed to provide a comprehensive evaluation of the hybrid artificial intelligence models in the classification of mammogram images. The results of this study have the potential to contribute to the development of more accurate and efficient diagnostic tools in the field of medical imaging, as shown in Figure 10.

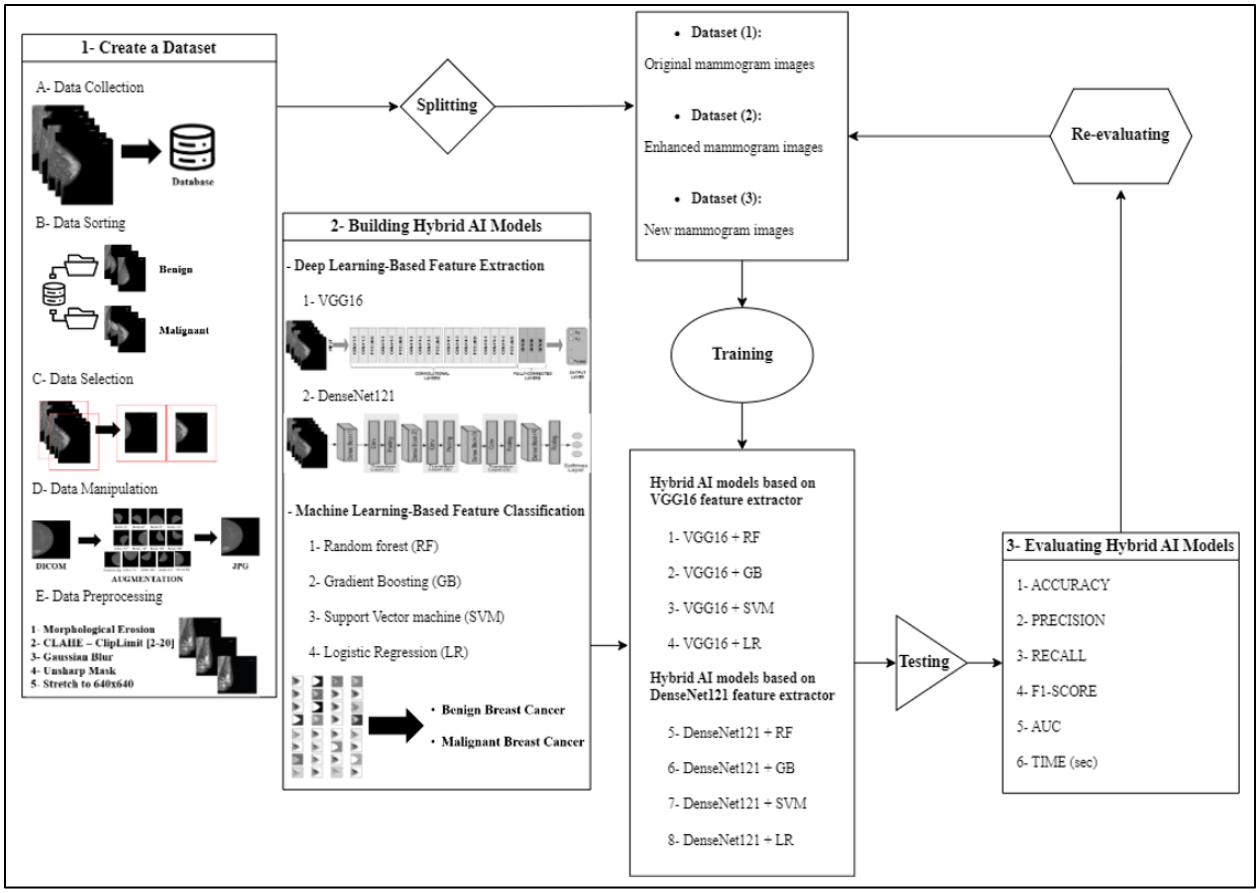

Figure 10: The methodological approach underpinning the research design employed in this study.

3.1 Ethical statement and confidentiality

Ethical approval was obtained from the Deanship of Scientific Research at Al-Quds University - Abu Dis, with the endorsement of the Institutional Review Board (IRB) and the ethical review committees of the participating hospitals where the data was collected. To protect patient privacy, all identifiable information was removed from the patient profiles, and strict measures were implemented to ensure the meticulous maintenance of confidentiality for the gathered data throughout the research process.

3.2 Data Collection

Initially, the research team collected all mammograms taken at the Dunya Women's Cancer Center between 2018 and 2023 from k-pacs workstation. This comprehensive dataset encompassed imaging data from 14,455 patients who had visited the center during this period. For each patient, two mammogram images were captured, one in the cranio-caudal (CC) position and the other in

the medio-lateral oblique (MLO) position, utilizing the MAMMOMAT Revelation device. In addition to the mammogram images, biopsy results from the same timeframe were also gathered. These biopsy records and results were collected from the Digisono 5V software, version 5.7.84, and included information about the affected side, amounted to a total of 4,976 cases, as shown in Table 5. To ensure proper organization and traceability, each patient was assigned a unique virtual serial number. Subsequently, the corresponding mammogram images for each patient were meticulously saved in individual files, with each file bearing the patient's specific serial number for easy identification and retrieval. Furthermore, the biopsy results, along with details about the affected side, were diligently recorded in an Excel sheet, with each entry linked to the respective patient's unique serial number. This meticulous documentation process facilitated the seamless correlation of imaging data with biopsy results during subsequent analysis. To safeguard the integrity and accessibility of the collected data, all the information, including the mammogram images, biopsy records, and associated Excel sheets, was securely downloaded onto an external hard disk for long-term preservation and future research endeavors.

Table 5: The distribution of the number of cases and biopsies gathered between 2018 and 2023 for constructing the database.

Date	Number of mammogram cases	Number of tru-cut biopsy cases
January – December (2018)	2,133	633
January – December (2019)	2,689	782
January – December (2020)	2,536	670
January – December (2021)	2,313	889
January – December (2022)	3,004	1,016
January – August (2023)	1,780	986
Total	14,455	4,976

3.3 Mammography Equipment and Tru-Cut Biopsy Procedure

The Dunya Women's Cancer Center employs the state-of-the-art mammomat Revelation device for conducting high-precision mammograms, playing a crucial role in the early detection and diagnosis of breast cancer. The Mammomat Revelation was introduced in 2014 as Siemens Healthineers' latest digital mammography system, as shown in Figure 11. It features a 50-degree wide-angle tomosynthesis capability that allows radiologists to see more of the breast tissue in 3D with improved depth resolution compared to earlier narrow-angle systems. The two most common breast positions used with the Mammomat Revelation are craniocaudal (CC) and mediolateral oblique (MLO) views. In the CC view, an image is taken from top to bottom of the breast when the breast is compressed between the compression paddle and the detector. In the MLO view, an image is taken from the side of the breast at a 45-degree angle when the breast is similarly compressed. As for technique factors, the Mammomat Revelation allows selection of mAs (milliampere-seconds) values between 10-715 mAs and kVp (kilovoltage peak) values between 23-49 kVp. The standard mammography technique is around 26-32 kVp and 50-200 mAs depending on breast thickness and density, as shown in Table 6. The system automatically selects the appropriate technique using its Automatic Exposure Control (AEC) based on breast composition and thickness (Vancoillie et al., 2021).

Table 6: Mammomat Revelation device components and features (Healthineers, 2023).

MAMMOMAT Revelation	**Parameters**
System specifications	
Basic unit	
Source-image distance	65 cm (25.6 Inch), for high geometric resolution and optimum patient access during positioning
Compression	3 kg (6.5 Ibs) to 20 kg (44 Ibs), automatic (OpComp) and manual adjustment
Collimation	Automatic for all sizes
Grid	Reciprocating, grid ratio 5:1, 31 lines/cm
X-ray generator	
Power output	5 kW (30 kV, 1 s, 60 s cycle time, acc. to IEC 60601-2-45)

kV range	23 kV to 35 kV (adjustable in 1 kV increments) 45 kV to 49 kV (adjustable in 1 kV increments)	
mAs range (at 25 Kv and maximum power)	With large focal spot - 2 mAs to 630 mAs manual mode - 7 mAs to 715 mAs in AEC mode	With small focal spot - 2 mAs to 200 mAs manual mode - 7 mAs to 240 mAs in AEC mode
X-ray tube unit		
Focal spot nominal value	Tungsten focal spot: 0.15 (small) / 0.3 (large) (IEC 60336)	
Anode-filter combinations	W/Rh, W/Ti	
Flat detector (Solid-state detector of amorphous selenium (aSe))		
Detector size	24 cm x 30 cm (9.5 Inch x 12 Inch)	
Material	Amorphous selenium (aSe)	
Conversion	direct-to-digital	
Pixel size	85 µm x 85 µm squared	
Image matrix	2816 x 3584 (24 cm x 30 cm / 9.5 Inch x 12 Inch)	
Breast Tomosynthesis Option		
Tube angulation	± 25°	
Scan time	< 25 seconds	
Number of projections	25	
Pixel size tomosynthesis	85 µm	
Distance between reconstructed slices	1 mm	
Reconstruction algorithms	Analytical or a unique combination of iterative and machine learning algorithms for EMPIRE	

Figure 11: An image showcasing the Mammomat Revelation device (Healthineers, 2023).

To confirm the nature of the findings from the mammogram, a biopsy is employed as the gold standard for diagnosing breast cancer and determining the specific characteristics of any detected abnormalities. This is the method approved by the Dunya Women's Cancer Center, ensuring the highest standard of care and accuracy in the diagnostic process. A tru-cut biopsy is a type of core needle biopsy used to sample tissue from the body for diagnostic testing, as shown in Figure 12. A tru-cut biopsy uses a hollow needle with a sharp cutting edge to remove a small cylinder of tissue from organs, lymph nodes, or masses. The needle is attached to a handle and works like a corkscrew. As the needle is inserted into the target area, a cutting edge slices off a "core" of tissue and retains it inside the hollow center of the needle as it is withdrawn (Minkowitz et al., 1986). The procedure is usually done under imaging guidance like ultrasound or CT scan so the doctor can see the target area in real-time and guide the needle precisely. This helps ensure the sample is taken from the correct location. Once the tissue core is collected, it is removed from the needle and sent to the lab for examination under a microscope by a pathologist. The pathologist can analyze the tissue architecture, look for abnormal cells or signs of infection/inflammation, and provide a diagnosis. For example, they may be able to tell if a mass is benign or cancerous. Tru-cut biopsies are considered a low-risk procedure and provide doctors with diagnostic tissue without requiring surgery. The sample allows for an accurate diagnosis to determine the best treatment options. The small incision also heals quickly with little scarring (Günes, 2018).

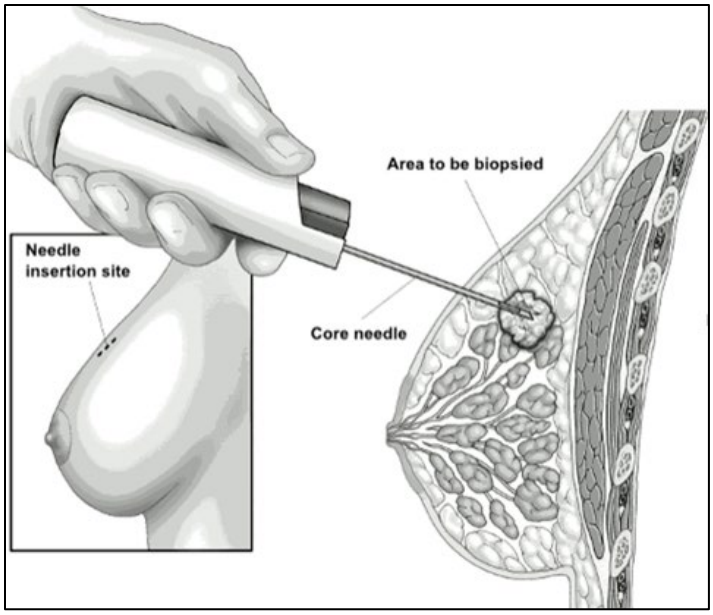

Figure 12: The procedure for obtaining a sample using the tru-cut technique (TEAM, 2018).

3.4 Data Sorting

The process of data sorting in this research involved several meticulous steps to ensure the accuracy and reliability of the dataset. After using Python to remove patient-identifying information from the image frames to protect privacy, a doctor with expertise in the field carefully reexamined and categorized the biopsy results into three distinct groups: normal, benign, and malignant, as shown in Table 7. This step was crucial in establishing the ground truth for the dataset and forming the basis for the subsequent sorting of mammogram images. The collection of mammogram images was then meticulously sorted into these categories, taking into account the affected side. This careful categorization was essential to ensure that the dataset accurately represented the diversity of cases and pathological conditions. Additionally, the strict criteria used in this research led to the exclusion of certain images, a necessary step to maintain the reliability and accuracy of the database.

Upon reviewing the number of images in each category, it became evident that there was a significant imbalance, which could potentially introduce bias in subsequent analyses and model training. As a result, deliberate efforts were made to balance the number of images in each category, aiming to prevent any potential bias and ensure the fairness and reliability of the dataset for future analysis and model training. The comprehensive and meticulous data sorting process, from categorizing biopsy results to organizing mammogram images, was fundamental in creating

a reliable and accurate dataset for the research, laying a strong foundation for subsequent analysis and model development.

Table 7: The distribution of the number of cases and images following classification by a radiologist, based on the biopsy reports associated with each case.

Category	Normal (number of images)	Known Biopsy-Proven (Benign) (number of images)	Known Biopsy-Proven (Malignant) (number of images)	Total
Number of cases	1,703 (3,410)	1,684 (3,371)	1,589(3,342)	4,976 (10,123)

3.5 Inclusion and Exclusion Criteria for Collected Mammograms

In this research, the inclusion criteria were precisely determined to ensure that only high-quality data models were included in the process of building the hybrid artificial intelligence model for breast cancer detection and classification. These criteria were established to select data models that are free of defects and problems, thus aiming to achieve the highest possible diagnostic accuracy. The inclusion criteria encompass patients who underwent mammography at the Dunya Women's Cancer Center in Palestine between 2018 and 2023, where the presence of breast cancer was suspected. Subsequently, a biopsy was conducted within the same timeframe to validate the findings of the mammogram. Upon confirmation of cancer following the biopsy results, these patients are considered eligible for inclusion in the study. Additionally, only cases where the biopsy results were issued and the patient was definitively diagnosed with breast cancer during the specified period are included.

Conversely, the exclusion criteria were also carefully defined to exclude any data models that did not meet the required standards for accuracy and reliability. This meticulous approach was taken to ensure the robustness and effectiveness of the hybrid artificial intelligence model in its application to breast cancer detection and classification. Exclusion criteria include the following:

1. Mammograms performed outside of Palestine, patients who had a mammogram and biopsy outside the specified period from 2018 to 2023, and cases where the mammogram exhibits artifacts, including motion, metallic, or any other type of artifact, can be combined into one

point: Mammograms or cases that were performed or occurred outside the specified location, time period, or had technical issues affecting the quality of the mammogram.

2. Patients undergoing mammography for follow-up purposes, patients who had a mammogram after undergoing treatments such as chemotherapy, radiation, lumpectomy, or mastectomy, and patients whose mammogram and biopsy results indicate a disease other than breast cancer, such as axillary lymph nodes or other conditions, can be combined into one point: Mammograms or cases related to follow-up, post-treatment monitoring, or non-breast cancer conditions.
3. Patients who refused to undergo biopsy or had it performed at a different facility, instances where the affected side, or both, was not clearly identified in the mammogram or pathology report, cases where there are issues with the pathology report due to sample problems or insufficiency, and patients for whom the center declined to provide data due to privacy concerns or the patient's wishes, can be combined into one point: Incomplete or missing data due to patient refusal, procedure performed elsewhere, inadequate documentation, or privacy concerns.

3.6 Data Manipulation

In this research, data augmentation played a pivotal role in expanding the dataset and ensuring its robustness for training machine learning models. The initial dataset, consisting of 10,123 images categorized into normal, benign, and malignant classes, was deemed insufficient to fully capture the diversity of features necessary for accurate model training. To address this limitation, a data augmentation technique was employed. The data augmentation process involved a series of image transformations, including rotation at multiple angles (90º, 180º, and 270º), adjustments in width and height, rescaling, shearing, zooming, horizontal flip operations, and specific filling techniques, as shown in Figure 13 and Table 8 (Goceri, 2023). These transformations effectively increased the dataset by 4 times, resulting in approximately 40,492 images, with 13,000 representing each of the normal, benign, and malignant categories by Python environment (Anaconda3). By significantly expanding the dataset through data augmentation, the research aimed to prevent overfitting and ensure that the machine learning model could effectively learn from a diverse range of image variations, ultimately leading to improved accuracy and generalization.

Table 8: Types of data and augmentation parameter.

Data Augmentation Types	Zoom	Rotation	Width shift	Height shift	Shear	Horizontal flip	Vertical flip
Parameters	0.2	5 degrees	0.15	0.15	0.01	True	True

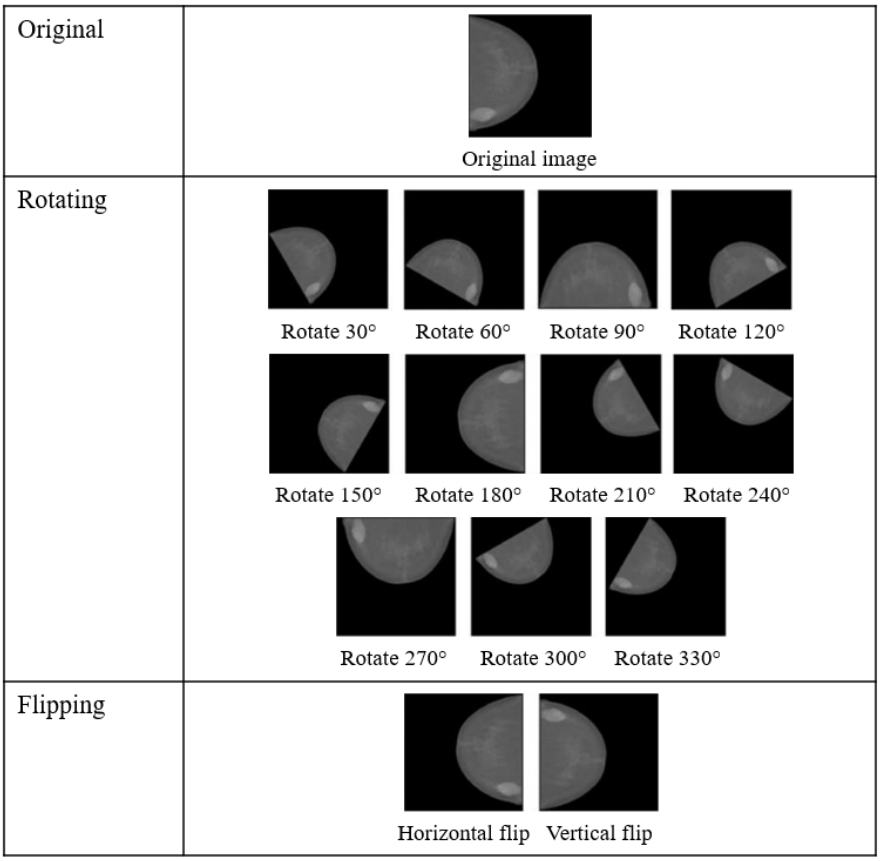

Figure 13: The outcomes obtained from employing augmentation methods on mammogram images, encompassing rotations and flips techniques.

After performing a multiplication process for the images, a conversion process was then executed to transform all the resulting images from dicom format to jpeg format. This conversion was carried out to facilitate the utilization of the images in the study. Converting the images to jpeg format offered several benefits. Firstly, it significantly reduced the file sizes, optimizing storage and making it more efficient to process and transfer the images. This was particularly advantageous in handling a large volume of medical images, as it minimized the storage requirements and accelerated data processing. Moreover, the jpeg format is widely supported across various software

and platforms, enhancing the accessibility and usability of the images for analysis and model training. This interoperability allowed for seamless integration of the images into the study's analytical framework and machine learning model development. Additionally, the conversion to jpeg format simplified the visualization and sharing of the images, enabling easier collaboration and communication among researchers and medical professionals involved in the study.

3.7 Data Preprocessing

The study's approach involves systematically implementing a sequence of advanced preprocessing techniques to enhance mammography images for analysis using deep and machine learning models. The morphological erosion preprocessing technique was utilized as the initial step in the image enhancement process. This technique aimed to reduce the complexity and granular noise present in raw mammographic images, thereby enhancing the visibility of significant structures within the breast tissue and suppressing irrelevant noise (Makandar & Halalli, 2015). Additionally, the technique involved the convolution of a specific structuring element across the image. This carefully defined matrix, in terms of size and shape through preliminary experiments, interacted with the local pixels of the image. When the pixel configuration did not match the structuring element, the process would adjust the central pixel, gradually decrease the boundaries, and lower the dimensionality of the structures (Makandar & Halalli, 2015).

In the second step, the Contrast-Limited Adaptive Histogram Equalization (CLAHE) technique was employed (Kharel et al., 2017). CLAHE is an image processing technique that improves the contrast of images by redistributing the intensity values of the pixels. It is a variant of adaptive histogram equalization (AHE), which also redistributes intensity values but can lead to over-amplification of noise in relatively homogeneous regions of an image. In this phase, the mammography image is segmented into distinct tiles. Each tile undergoes individual histogram equalization(Dabass et al., 2019; Kharel et al., 2017). CLAHE's primary strength is its ability to highlight local contrast, enhancing sophisticated details like microcalcifications. An essential aspect of this method is the inclusion of a contrast threshold. If a histogram bin exceeds this threshold, the excess is spread out to the neighboring bins. This helps avoid excessive contrast boosts that could amplify noise. CLAHE mitigates this concern by controlling the contrast boost in different image sections. It achieves this by limiting the histogram of every segment to a set

value before determining the cumulative distribution function (CDF). The CLAHE estimated according to the following:

$$CDF(x) = \sum_{y=0}^{x} \frac{H(y)}{NT} \quad \textbf{(1)}$$

Where:

- CDF(x) is the cumulative distribution function at intensity level x.
- H(y) is the histogram value at intensity level y.
- N is the total number of pixels in the image.
- T is the clipping value.

The clipping value determines the balance between contrast enhancement and noise reduction. A higher value reduces contrast but cuts noise more, while a lower one increases contrast but reduces noise less. To address noise issues in consistent areas of an image, CLAHE applies:

- Histogram clipping: Each region's histogram is controlled at a set value before determining the CDF, reducing excessive noise amplification.
- Larger tile size: The image is segmented into grid tiles, each treated separately, reducing noise's effect on enhancing contrast.
- Bilinear interpolation: When merging processed tiles for the final image, CLAHE employs bilinear interpolation, smoothing tile edges and lessening noise-induced artifacts.

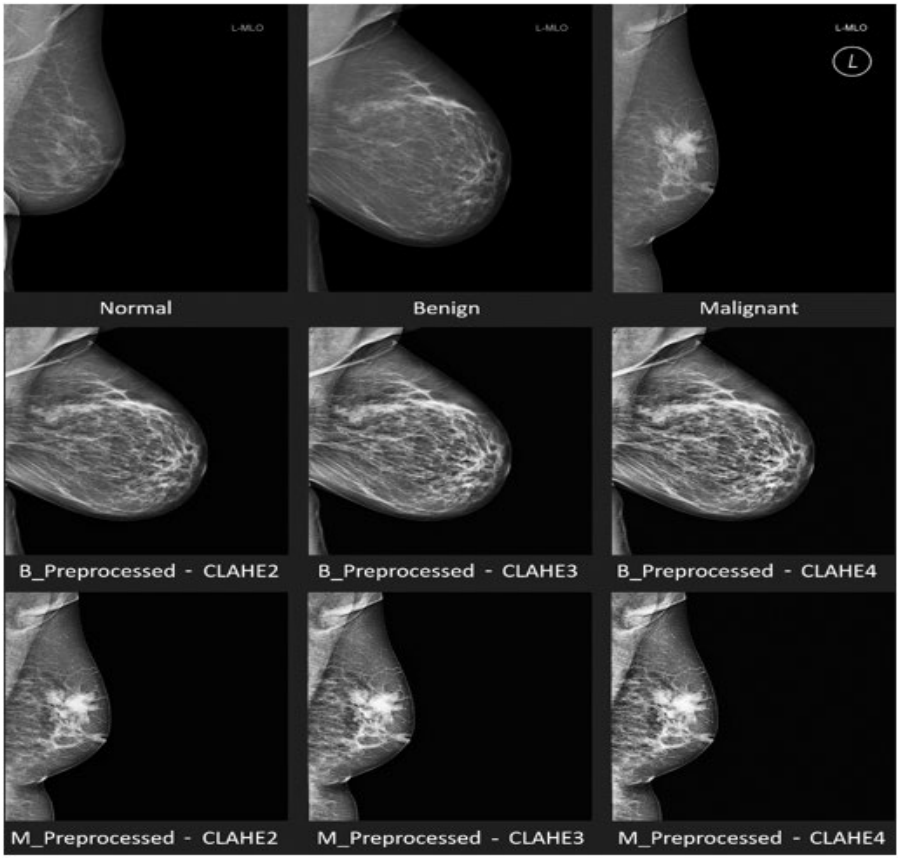

Figure 14: The effectiveness of CLAHE algorithm in enhancing benign and malignant mammography images at different clip limits (0, 2, 3, and 4).

Figure 14 illustrates the effectiveness of CLAHE in enhancing mammography images, showcasing the improvements in image contrast and quality achieved through this image processing technique.

Furthermore, the Effective Measure of Enhancement (EME) was used to quantify and measure the image enhancement obtained by splitting the image into several blocks. The following equation is utilized for EME:

$$EME = \frac{1}{K_1 K_2} \sum_{L=1}^{K_2} \sum_{K=1}^{K_1} 20\, log \left(\frac{I_{max}(k,l)}{I_{min}(k,l)} \right) \quad \textbf{(2)}$$

Where K_1, K_2 are the number of horizontal and vertical blocks in the image, $I_{max}(k,l)$, and $I_{min}(k,l)$ are the maximum and minimum pixel values in each block.

The Peak Signal to Noise Ratio (PSNR) was used to measure the deviation of the current image from the original image with respect to the peak value of the gray level. Given a reference image f and a test image$g$, both of size$M \times N$, the PSNR between f and g is defined by:

$$PSNR(f,g) = 10\,log_{10}\left(\frac{255^2}{MSE(f,g)}\right) \quad \textbf{(3)}$$

Where,

$$MSE(f,g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(f_{ij} - g_{ij}\right)^2 \quad \textbf{(4)}$$

As the Mean Squared Error (MSE) approaches zero, the PSNR value tends to approach infinity. This relationship demonstrates that a higher PSNR value corresponds to a higher image quality. In other words, as the PSNR value increases, the image quality improves, indicating a closer resemblance between the original and the enhanced image (Kharel et al., 2017).

In the third step, the CLAHE enhanced images were subjected to the Laplacian of Gaussian (LoG) edge enhancement (Maitra et al., 2012). It is a two-steps procedure commences with the image being mildly smoothed using a Gaussian filter to reduce sensitivity to noise. Thereafter, a Laplacian operator identifies areas of rapid intensity transition, indicative of edges. This integration of Gaussian and Laplacian techniques improves the delineation of pathological structures. In the final step, the Unsharp Masking technique, a digital image sharpening method, has been utilized to enhance the clarity of details in mammography images, an essential step in improving diagnostic accuracy in our study (Duan et al., 2018). This technique primarily works by amplifying the contrast at the edges within an image, thereby augmenting the sharpness and making the finer details more visible.

In the implementation of Unsharp Masking, our first step involved creating a slightly blurred version of the original mammographic image, achieved by applying a Gaussian blur filter. This process effectively suppressed noise and reduced high-frequency details, creating what is referred to as the "unsharp" image. Subsequently, this blurred version was subtracted from the original image, resulting in the creation of a high-contrast mask that represented the detailed components removed during the blurring process. This high-contrast mask was then carefully added back to the original mammography image. The purpose of this step was to ensure that the reintroduction of the details occurred with an emphasis on their clarity, enhancing the sharpness of structures and features within the breast tissue. We precisely controlled the degree of sharpening by adjusting the extent to which the high-contrast mask was combined with the original image. This adjustment was vital to avoid over-sharpening that might introduce artificial features or noise, potentially

leading to misinterpretation. Having experienced this comprehensive enhancement, the images are then presented to the deep convolutional neural networks (DCNN) for analysis. Trained on such enhanced images, the model examines and classifies potential pathologies using machine learning algorithms. Radiologists can then employ these clearer, enhanced images to augment their diagnostic accuracy. The structured integration of these methods ensures the comprehensive and optimal processing of each mammography image, aiming to enhance the precision of deep learning evaluations and, by extension, the final diagnostic outcomes.

3.8 Dataset

The supervised dataset used in this research consists of 40,492 images that have been enhanced using Contrast Limited Adaptive Histogram Equalization (CLAHE) and saved in JPEG format with a resolution of 640x640 pixels. This dataset was specifically prepared to facilitate the development of a hybrid model that integrates deep learning and machine learning techniques for the analysis of mammogram images. To effectively utilize this dataset, it was divided into two separate databases. The first database contained images categorized as normal and benign, while the second database included images categorized as normal and malignant, following a Binary database system. This categorization allowed for the creation of distinct subsets tailored to the specific characteristics of the images, enabling focused analysis and model training. Subsequently, the dataset was further partitioned to serve different purposes. Approximately 70% of the dataset was allocated for training the model, enabling it to learn and adapt to the patterns and features present in the images. The next 20% of the dataset was reserved for testing the model's performance, providing an independent set of images to evaluate its accuracy and generalization. Finally, 10% of the dataset was designated for verification, serving as a separate subset to validate the model's robustness and reliability. This meticulous approach to dataset creation and partitioning ensures that the dataset is effectively utilized for training, testing, and validating the hybrid model, ultimately contributing to the model's accuracy and effectiveness in analyzing mammogram images.

Cross-validation is a crucial technique used in machine learning and model training to assess the performance and generalization of a predictive model. It involves partitioning the dataset into complementary subsets, performing the analysis on one subset (training set), and validating the analysis on the other subset (testing set). This process is repeated multiple times, with each subset

used as the testing set exactly once. The primary goal of cross-validation is to evaluate how the results of a statistical analysis will generalize to an independent dataset. It helps to assess how well a model performs on unseen data and provides an estimate of the model's predictive performance. Common methods of cross-validation include k-fold cross-validation, where the dataset is divided into k subsets and the model is trained and tested k times, and leave-one-out cross-validation, where each data point is used as the testing set in turn, with the rest of the data used for training. By using cross-validation, machine learning practitioners can gain insights into how well their models generalize to new data, identify potential issues such as overfitting, and make informed decisions about model selection and hyperparameter tuning. For this investigation, cross-validation was employed to train artificial intelligence models. The entire dataset was divided into 25 folds and splits, as shown in Figure 15. The dataset used in this research consists of three types of data collections:

1. Non-enhanced Mammogram Images: This subset contains 40,492 original mammography images in their raw form, without any enhancement or preprocessing applied.
2. Enhanced Mammogram Images: This subset includes the same 40,492 mammography images, but they have undergone Contrast Limited Adaptive Histogram Equalization (CLAHE) preprocessing. CLAHE enhances the contrast and visibility of structures within the images.
3. New Enhanced Mammogram Images: This subset consists of additional mammography images that were acquired separately and then enhanced using the same CLAHE technique.

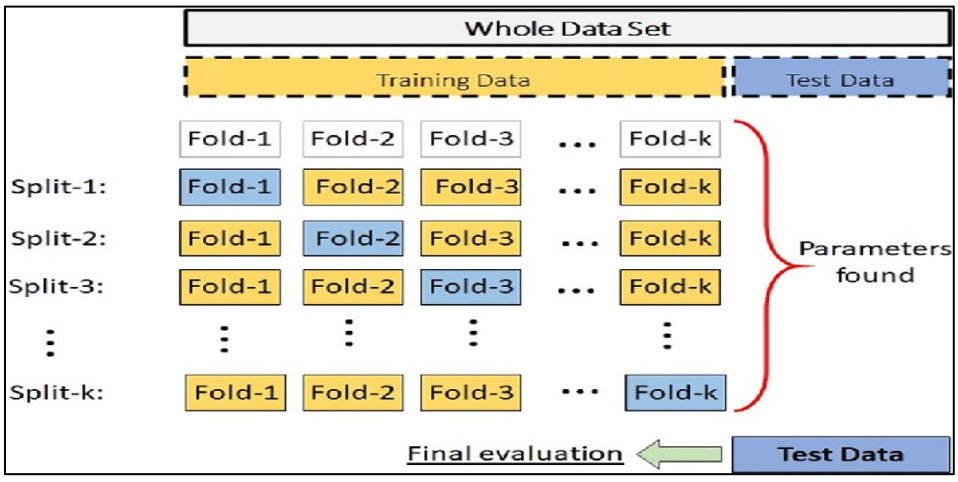

Figure 15: The process of dividing data into separate subsets for cross-validation, which are used for training and evaluating artificial intelligence models (Duran-Lopez et al., 2020).

3.9 Building a Hybrid Artificial Intelligence (AI) Model

In this section of the research methodology, the focus is on a critical stage of the study, which involves the development of a hybrid artificial intelligence model aimed at detecting and classifying breast cancer from mammogram images. This hybrid approach comprises two integral components that complement each other. The first component utilizes deep learning, particularly Deep Convolutional Neural Networks (DCNN), to extract features from mammogram images. These extracted features are then employed for the classification of benign and malignant cancer types using machine learning algorithms. Subsequently, these models undergo rigorous training, testing, and validation to ensure their efficacy for diagnostic purposes in the context of breast cancer.

3.9.1 Deep Learning-Based Feature Extraction

In this research, the decision to employ deep learning for feature extraction from mammogram images stemmed from its capacity to automatically discern and extract intricate patterns and features, which are crucial for the accurate detection and classification of breast cancer indicators. Deep learning, as opposed to traditional machine learning methods, has demonstrated superior performance in handling complex and unstructured data, making it an ideal choice for analyzing mammogram images, which often contain subtle and nuanced features indicative of breast cancer. The Python programming language, renowned for its versatility and extensive libraries tailored for machine learning and deep learning tasks, was utilized within the Anaconda environment to develop and implement the hybrid model. This environment provided a comprehensive suite of tools and packages, facilitating the seamless integration of various deep learning frameworks and libraries, such as TensorFlow, Keras, and PyTorch, essential for constructing and training the deep learning model.

Specifically, the Deep Convolutional Neural Network (DCNN) was selected as one of the primary deep learning approaches in this study. DCNNs are well-suited for image recognition tasks, as they are adept at automatically learning hierarchical representations of visual data, making them particularly effective for feature extraction from mammogram images. By leveraging the hierarchical structure of DCNNs, the model was able to capture and analyze intricate patterns and textures within the mammogram images, enabling the automatic extraction of relevant features crucial for the accurate identification of potential abnormalities associated with breast cancer.

DCNNs encompass a diverse range of architectures, each with unique characteristics and capabilities. Notable DCNN architectures include AlexNet, VGG (e.g., VGG16, VGG19), GoogLeNet (Inception), ResNet, and DenseNet, among others. These architectures vary in terms of depth, connectivity, and design principles, influencing their performance in feature extraction and representation learning tasks. AlexNet, a pioneering DCNN architecture, introduced the concept of deep convolutional neural networks and demonstrated remarkable performance in image classification tasks. VGG, characterized by its uniform architecture with multiple convolutional layers, has been widely recognized for its ability to learn discriminative features from images. GoogLeNet, known for its Inception modules, excels in capturing diverse and multi-scale features within images. ResNet, renowned for its residual connections, addresses the challenge of training very deep networks and facilitates the learning of highly abstract features. DenseNet, distinguished by its densely connected layers, promotes feature reuse and facilitates the flow of information throughout the network, leading to efficient feature extraction and representation learning.

In the context of this study, the specific focus is on VGG16 and DenseNet 121. VGG16, with its deep architecture comprising 16 weight layers, excels in capturing detailed and abstract features from images, making it well-suited for tasks requiring fine-grained feature extraction. On the other hand, DenseNet 121, with its densely connected layers, fosters enhanced feature reuse and facilitates the flow of information throughout the network, potentially enhancing its ability to capture nuanced and interrelated features within the mammogram images. By comparing the feature extraction capabilities of VGG16 and DenseNet 121, my study aims to evaluate their performance in discerning subtle abnormalities indicative of breast cancer. This evaluation will encompass metrics such as feature representation quality, computational efficiency, and overall classification accuracy, ultimately identifying the most effective model for extracting features relevant to the accurate detection and classification of breast cancer indicators from mammogram images. VGG16 is a convolutional neural network model that was developed by researchers at the University of Oxford's Visual Geometry Group. This model is well-known for setting the precedent in 2014 on using very deep convolutional networks for large-scale image classification. When applied to mammogram images for breast cancer detection and diagnosis, VGG16 has been shown to extract a variety of useful features from the images at different levels of abstraction, as shown in Figure 16 (Sivanantham et al., 2023).

At the lowest level, VGG16 analyzes mammograms to extract basic texture and edge features. This allows it to identify low-level patterns within the images like lines, blobs and other simple shapes that may indicate abnormalities such as masses, microcalcifications, architectural distortions, etc. Mid-level features captured by VGG16 provide information on the morphological characteristics of any lesions detected. Features like size, density, circularity and other metrics help characterize lesions as benign or malignant. VGG16 is also able to extract high-level semantic features from mammograms through its deep convolutional layers. These features allow it to recognize more complex patterns and classify different areas of the breast or overall mammogram images into diagnostic categories (Liu et al., 2024). By utilizing a hierarchical feature extraction approach, VGG16 analyzes mammograms at both local and global levels. The multi-scale feature maps generated by VGG16 at various depths of the network allow it to effectively differentiate between normal and abnormal findings for the task of breast cancer detection. The features extracted from pre-trained VGG16 models have also been used as inputs to other classification algorithms, helping to further improve their diagnostic performance on mammogram images. Overall, VGG16 has proven to be well-suited for automated analysis of mammograms due to the diverse set of features it can extract from the images (Sivanantham et al., 2023).

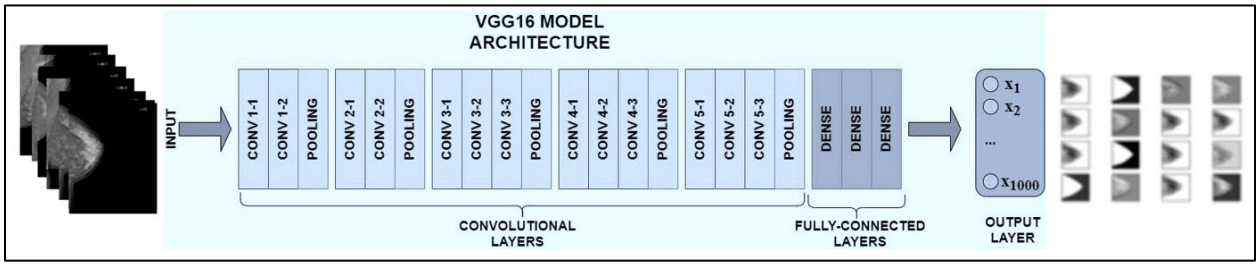

Figure 16: VGG16 model architecture.

In comparison, DenseNet 121, a convolutional neural network architecture, is proficient at extracting a diverse array of features from mammogram images, which are pivotal for the detection and diagnosis of breast cancer, as shown in Figure 17. One crucial set of features it captures is texture features, encompassing the distribution of microcalcifications, architectural distortions, and masses within the breast tissue. These texture patterns are vital for identifying abnormal tissue structures that may indicate the presence of cancer. By analyzing these texture features, DenseNet 121 can assist in identifying potential areas of concern within the mammogram images. Moreover, DenseNet 121 can extract shape features from mammogram images, including the size, contour, and spatial arrangement of potential lesions. These features play a significant role in characterizing

the morphology of suspicious areas within the breast tissue, aiding in the identification of potential malignancies. The network's ability to capture these shape features contributes to a more comprehensive analysis of the mammogram images, providing valuable insights into the structural characteristics of any detected abnormalities (Pattanaik et al., 2022).

The network's capability to capture margin features is also noteworthy. It can analyze the margins or boundaries of lesions within mammogram images, which is crucial as irregular or spiculated margins are often associated with malignant tumors. This capability assists in distinguishing between benign and malignant lesions, contributing to more accurate diagnoses. By extracting these margin features, DenseNet 121 can provide valuable information about the spatial characteristics of potential abnormalities, aiding in the assessment of their potential malignancy. Additionally, DenseNet 121 is capable of analyzing the density of breast tissue in mammogram images, an important factor in breast cancer detection. Dense breast tissue can make it more challenging to detect abnormalities, and the network's extraction of density features helps in this process (Rybiałek & Jeleń, 2020). By considering the density of the breast tissue, DenseNet 121 can provide valuable insights into the composition of the breast tissue, which is essential for a comprehensive analysis of the mammogram images. Lastly, the network considers the contextual information surrounding potential lesions in mammogram images, providing valuable insights into the spatial relationships between different structures within the breast tissue. This holistic approach to feature extraction enhances the network's ability to aid radiologists and clinicians in the early detection and diagnosis of breast cancer. By considering the contextual information, DenseNet 121 can provide a more comprehensive analysis of the mammogram images, aiding in the identification and characterization of potential abnormalities.

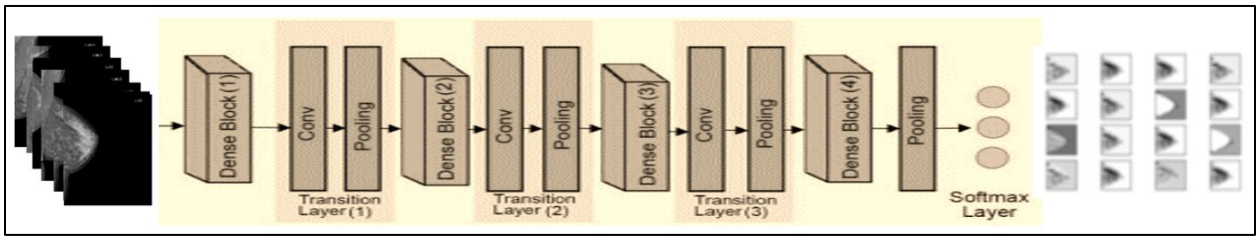

Figure 17: DenseNet121 model architecture.

3.9.2 Machine Learning-Based Feature Classification

Following the extraction of features from mammogram images using VGG16 or DenseNet121, a machine learning model was trained to classify breast cancer into benign or malignant cases. The

features extracted from the mammogram images, such as texture, shape, and other visual patterns, were used as input to the machine learning model. This allowed the model to learn and differentiate between benign and malignant cases based on these features. The decision to use machine learning for this classification task was based on previous studies that demonstrated the potential of machine learning in accurately classifying mammogram images. Machine learning has shown promise in its ability to effectively interpret complex image features, handle non-linear and high-dimensional data, generalize from training data to new cases, and build upon its success in medical imaging applications was pivotal. The intricate and informative features extracted from deep learning models like VGG16 and DenseNet121 necessitate sophisticated analysis, for which machine learning algorithms are well-suited. Additionally, the adaptability and generalization capabilities of machine learning models are essential for accurately classifying new, unseen cases, making them a natural fit for the task of classifying breast cancer based on mammogram features.

In this study, four distinct types of machine learning algorithms were carefully chosen based on comprehensive experimentation with various machine learning algorithms. The selected algorithms, namely random forest, gradient boosting, support vector machines (SVM), and logistic regression, were identified as top performers across multiple evaluation metrics, including accuracy, precision, recall, F1-score, area under the curve (AUC), and the time required for training and prediction. This rigorous selection process aimed to ensure that the chosen algorithms not only excelled in their predictive performance but also demonstrated efficiency in terms of computational resources and time. By considering a wide range of evaluation criteria, the study sought to identify the most suitable algorithms for the specific task of classifying breast cancer based on features extracted from mammogram images.

A. Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs for classification. The extracted features from VGG16 or DenseNet121 can be used as input to train the Random Forest classifier. The Random Forest algorithm will create multiple decision trees, each trained on a subset of the features and a subset of the training data, and the final classification is determined by aggregating the predictions from all the trees, as shown in Figure 18 (Rybiałek & Jeleń, 2020).

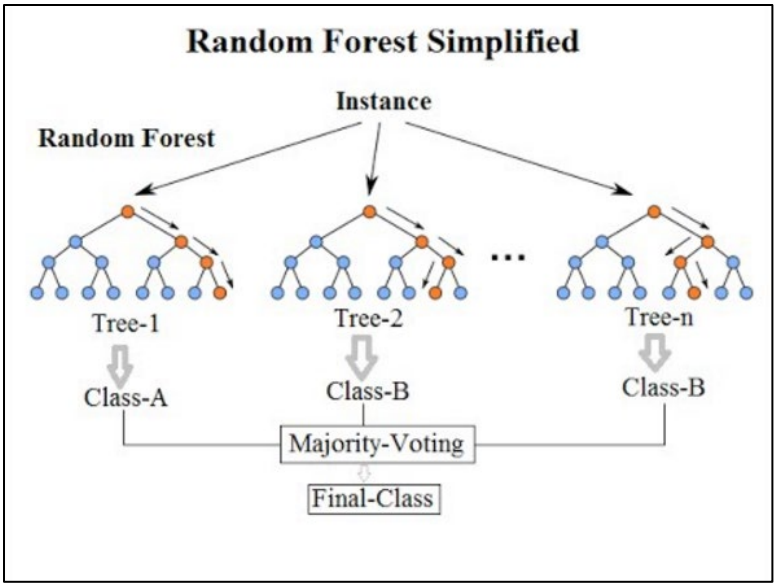

Figure 18: Random Forest (RF) (Rybiałek & Jeleń, 2020).

B. Gradient Boosting: Gradient Boosting is another ensemble learning technique that builds an additive model by sequentially training weak learners (e.g., decision trees) on the residual errors of the previous learners. The extracted features can be used as input to the Gradient Boosting classifier, which will iteratively train weak learners to minimize the classification errors made by the previous learners, as shown in Figure 19 (Rybiałek & Jeleń, 2020).

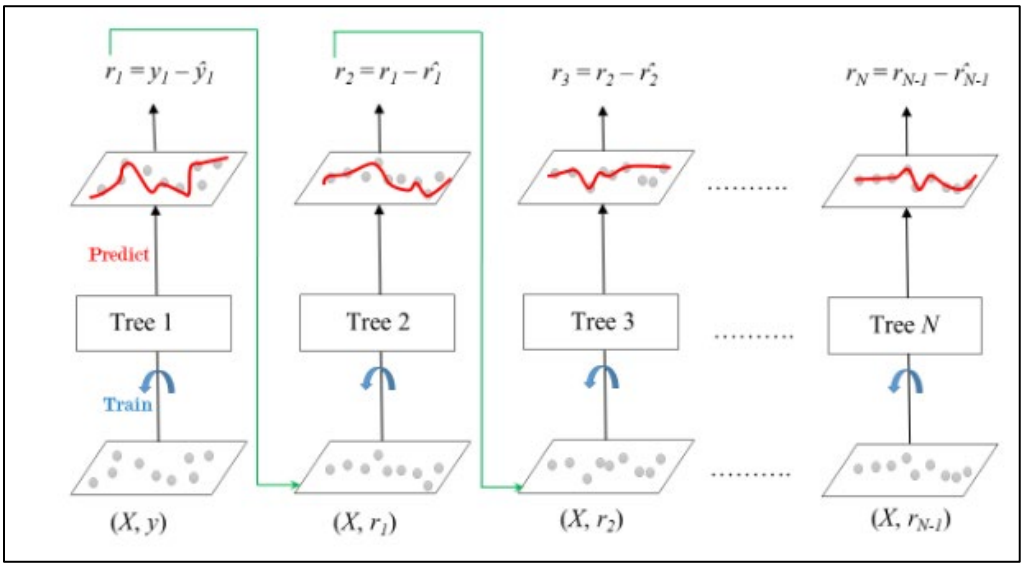

Figure 19: Gradient Boosting (GB) (Rybiałek & Jeleń, 2020).

C. Support Vector Machines (SVM): SVM is a supervised learning algorithm that finds the optimal hyperplane that maximizes the margin between classes in the feature space. The extracted features from VGG16 or DenseNet121 can be used as input to the SVM classifier, which will attempt to find the best decision boundary that separates the different classes in the feature space, as shown in Figure 20 (Azar & El-Said, 2014).

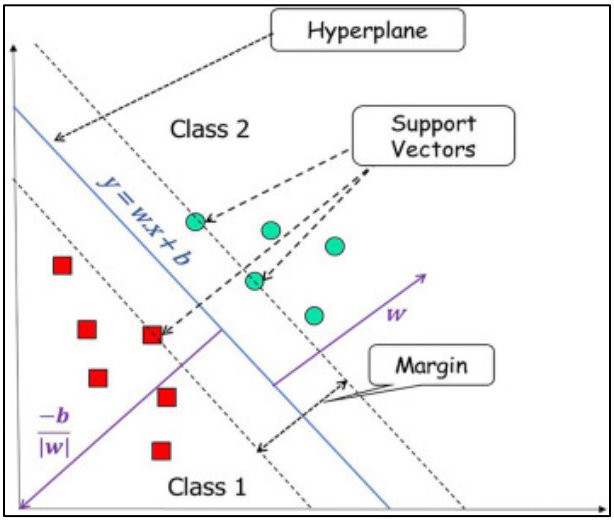

Figure 20: Support Vector Machine (SVM) (Azar & El-Said, 2014).

D. Logistic Regression: Logistic Regression is a statistical model that estimates the probability of an instance belonging to a particular class. The extracted features can be used as input to the Logistic Regression classifier, which will learn the weights or coefficients that best describe the relationship between the features and the class labels, as shown in Figure 21.

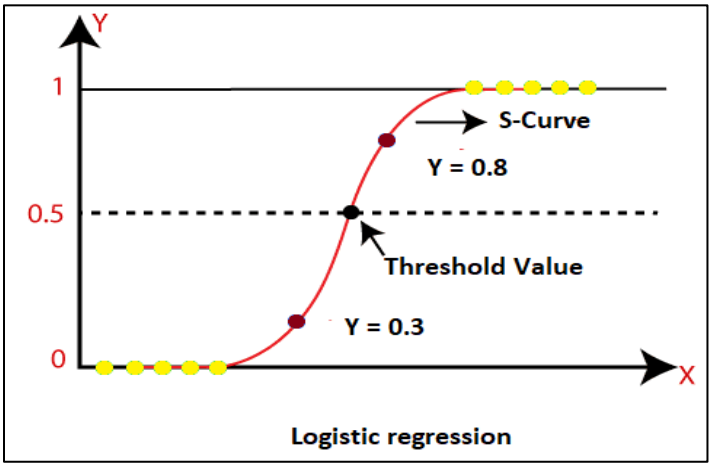

Figure 21: Logistic Regression (LR) (Ayer et al., 2010).

In all these cases, the extracted features from VGG16 or DenseNet121 serve as a high-level representation of the input images, capturing the relevant visual information that can be effectively utilized by the traditional machine learning classifiers. By leveraging the powerful feature extraction capabilities of these deep learning models, the traditional classifiers can potentially achieve better performance compared to using raw pixel data as input. To perform this operation, the system code was programmed in the Python language within the Anaconda Navigator (anaconda3) environment, using the JupyterLab interface as shown in Figure 22.

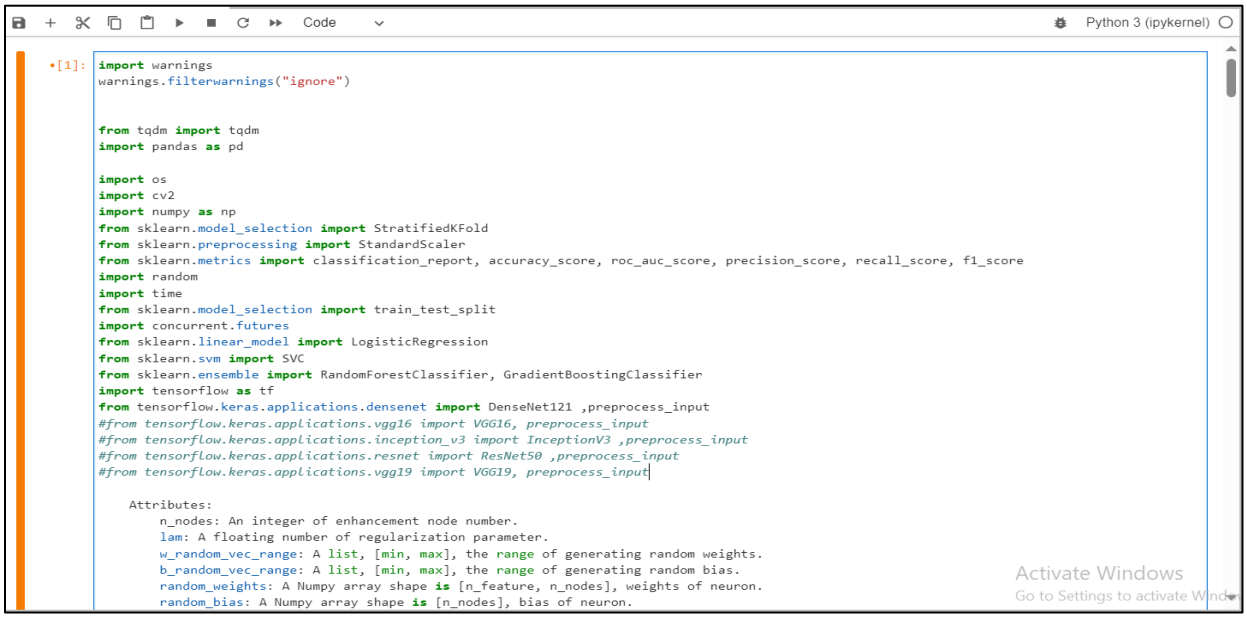

Figure 22: The Python environment leveraged to construct the code-base for the hybrid models investigated in this research, along with the various software libraries utilized.

3.10 Approaches for Evaluating a Hybrid Artificial Intelligence (AI) Model

Several tools are utilized to evaluate the performance of a classification model. The confusion matrix provides a detailed breakdown of the model's predictions and their correspondence with the actual outcomes. Accuracy measures the overall correctness of the model's predictions, while the receiver-operating curve (ROC) and the area under the ROC curve (AUC) are used to assess the model's ability to distinguish between classes. Precision quantifies the proportion of true positive predictions among all positive predictions, and the F1 score combines precision and recall into a single metric. The method employed in this study is widely recognized and commonly utilized in research within this field due to its established effectiveness and applicability.

3.10.1 The Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows visualization of the performance of an algorithm. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes. Usually, in the field of machine learning a confusion matrix is known as the error matrix, as shown in Fig 23. The matrix is a 2x2 table with four cells that represent the counts of the following:

1. True Positives (TP): The cases in which the model predicted "yes" and the actual value was also "yes."
2. True Negatives (TN): The cases in which the model predicted "no" and the actual value was also "no."
3. False Positives (FP): The cases in which the model predicted "yes" but the actual value was "no."
4. False Negatives (FN): The cases in which the model predicted "no" but the actual value was "yes."

The confusion matrix provides a more detailed understanding of how well the classification model is performing and can be used to calculate various performance metrics such as accuracy, precision, recall, and F1 score. Figure 28 provides an example of the confusion matrix for two class's classification (Beauxis-Aussalet & Hardman, 2014).

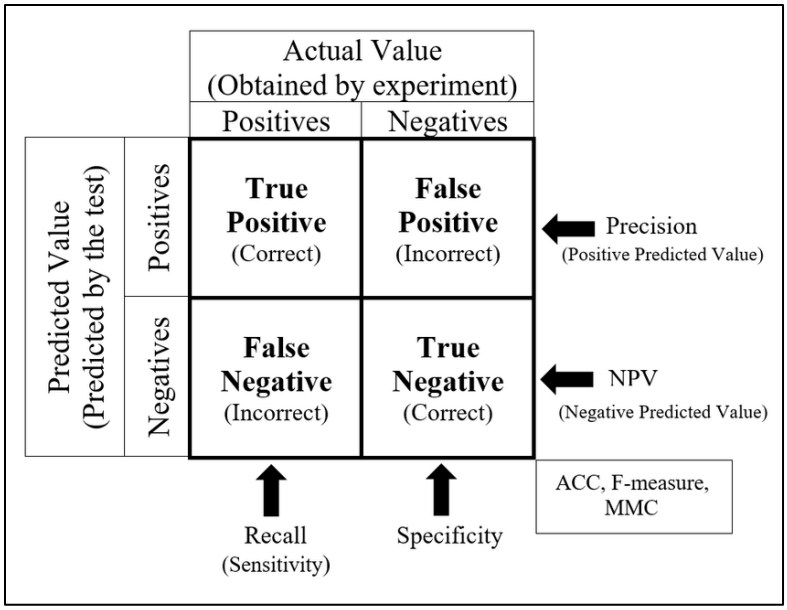

Figure 23: A Confusion Matrix, a tabular representation that illustrates the correlation between the actual values and the predicted values, highlighting the correct predictions as well as the errors in the predictive model. (Mokhtari et al., 2021)

3.10.2 The Accuracy

Accuracy is a fundamental metric used to evaluate the performance of a classification model. It measures the proportion of correctly classified instances out of the total instances evaluated. The accuracy of a model is calculated by dividing the number of correct predictions by the total number of predictions made. While accuracy is a simple and intuitive measure, it may not always provide a complete picture of a model's performance, especially when dealing with imbalanced datasets where one class is much more frequent than the other. In such cases, a high accuracy score can be misleading if the model is simply predicting the majority class for every instance. Therefore, while accuracy is a valuable metric, it is important to consider it alongside other performance measures such as precision, recall, and the F1 score, especially in scenarios where class imbalances exist. This comprehensive evaluation helps to gain a more nuanced understanding of the model's strengths and weaknesses (Hicks et al., 2022). The accuracy is defined as in equation 5.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(5)}$$

Where:

TP: True Positives, TN: True Negatives, FP: False Positives and FN: False Negatives.

3.10.3 The Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model across different threshold values, as shown in Figure 24. It illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) as the discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. A model with perfect discrimination has an ROC curve that passes through the upper-left corner of the plot, representing 100% true positive rate and 0% false positive rate. On the other hand, a model with no discrimination ability would have an ROC curve that is a 45-degree diagonal line from the bottom-left to the top-right of the plot. The area under the ROC curve (AUC) is also a commonly used metric to quantify the overall performance of a classification model. A higher AUC value indicates better overall performance, with a value of 0.5 representing a model with no discrimination ability and a value of 1 representing a model with perfect discrimination. The ROC curve and AUC are valuable tools for comparing the performance of different models and for selecting the optimal threshold for making predictions based on the specific needs of the application. The ROC analysis is a well-known evaluation method for detecting tasks. Firstly, a ROC analysis was used in medical decision-making; consequently, it was used in medical imaging (Hanley, 1989). They are defined as in equation 6-9.

$$True\ Positive\ Rate\ (TPR)(Sensitivity)(Recall) = \frac{TP}{FN + TP} = \frac{TP}{P} \quad \textbf{(6)}$$

$$False\ Positive\ Rate\ (FPR)(False\ Alarm\ Rate) = \frac{FP}{TN + FP} = \frac{FP}{N} \quad \textbf{(7)}$$

$$True\ Negative\ Rate(TNR)(Specificity) = \frac{TN}{TN + FP} = \frac{TN}{N} = 1 - FPR \quad \textbf{(8)}$$

$$False\ Negative\ Rate\ (FNR) = \frac{FN}{FN + TP} = \frac{FN}{P} \quad \textbf{(9)}$$

Where:

TP: True Positives, TN: True Negatives, FP: False Positives and FN: False Negatives.

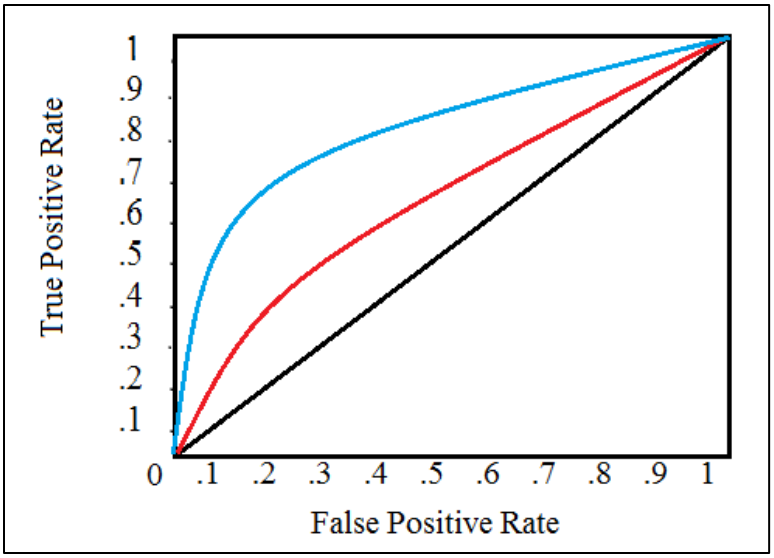

Figure 24: The Receiver Operating Characteristic (ROC) curve, a graphical plot that depicts the trade-off between the true positive rate and the false positive rate, providing a comprehensive evaluation of the forecasting model's performance (Soltani et al., 2019).

3.10.4 The Area under the ROC Curve (AUC)

The Area under the Receiver Operating Characteristic (ROC) curve, commonly referred to as AUC, is a widely used metric for evaluating the performance of a binary classification model. The AUC quantifies the model's ability to discriminate between positive and negative classes across all possible threshold values. The AUC is a value between 0 and 1, where a higher AUC indicates better overall performance. A model with an AUC of 0.5 has no discrimination ability and is essentially making random predictions, while a model with an AUC of 1 demonstrates perfect discrimination, achieving a 100% true positive rate and a 0% false positive rate. The AUC is particularly useful when dealing with imbalanced datasets, where one class is much more prevalent than the other. In such cases, accuracy can be misleading, but the AUC provides a more comprehensive assessment of the model's performance. Furthermore, the AUC is valuable for comparing different models and selecting the optimal model for a specific application. It provides a single scalar value that summarizes the model's performance across all possible classification thresholds, making it a powerful tool for model evaluation and selection. The AUC is used in the medical diagnosis system and it provides an approach for evaluating models based on the average of each point on the ROC curve (Sokolova et al., 2006).

3.10.5 The Precision

Precision is a crucial metric used to evaluate the performance of a classification model, particularly in binary classification tasks. It measures the proportion of true positive predictions out of all the instances that were predicted as positive by the model. Mathematically, precision is calculated as the ratio of true positive predictions to the sum of true positive and false positive predictions. In other words, it assesses the model's ability to accurately identify the relevant instances from the total instances it has predicted as positive. Precision is especially important in scenarios where the cost of false positives is high. For example, in medical diagnosis, precision is crucial because misclassifying a healthy individual as having a disease can have serious consequences. While precision is a valuable metric, it should be considered alongside other performance measures such as recall, accuracy, and the F1 score to gain a comprehensive understanding of the model's effectiveness. This holistic evaluation helps in making informed decisions about the model's performance in real-world applications. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low FPR (Tsopra et al., 2021). The precision is calculated using the following equation.

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(10)}$$

Where:

TP: True Positives and FP: False Positives.

3.10.6 The F1-Score

The F1 score is a metric that combines both precision and recall into a single value, providing a balanced assessment of a classification model's performance, particularly in binary classification tasks. The F1 score is calculated as the harmonic mean of precision and recall. It takes into account both false positives and false negatives, making it a useful measure when the class distribution is imbalanced. A high F1 score indicates that the model has both good precision and recall, striking a balance between minimizing false positives and false negatives. This makes the F1 score particularly valuable in situations where there is an uneven class distribution or when false positives and false negatives carry different costs. By considering both precision and recall, the F1 score provides a comprehensive evaluation of a model's ability to correctly identify positive instances while minimizing misclassifications. It is a widely used metric for comparing different

models and selecting the optimal model for specific applications. F1 score is the weighted average of precision and recall. It is used as a statistical measure to rate the performance of the classifier. Therefore, this score takes both false positives and false negatives into account (Chicco & Jurman, 2020). F1 score is defined as in Equation 11.

$$F1\ Score = \frac{2 \times Precisiqon \times Recall}{Precision + Recall} \quad \textbf{(11)}$$

3.10.7 The Recall

Recall, also known as sensitivity, is a crucial metric used to evaluate the performance of a classification model, particularly in binary classification tasks. It measures the proportion of actual positive instances that were correctly identified by the model. Mathematically, recall is calculated as the ratio of true positive predictions to the sum of true positive and false negative predictions. In essence, it assesses the model's ability to capture all positive instances from the total instances that are actually positive. Recall is especially important in scenarios where missing positive instances (false negatives) is more critical than including false positive predictions. For example, in medical diagnosis, recall is crucial because failing to identify a true positive case can have serious consequences. While recall is a valuable metric, it should be considered alongside other performance measures such as precision, accuracy, and the F1 score to gain a comprehensive understanding of the model's effectiveness. This holistic evaluation helps in making informed decisions about the model's performance in real-world applications (Müller & Braun, 2023).

3.10.8 The Time

The time metric evaluates the computational efficiency or inference speed of the model. It measures the time it takes for the model to process and generate predictions on a given set of input data, usually reported in seconds or milliseconds per image. This metric is important because faster inference times are desirable in real-world clinical applications, where radiologists and healthcare providers need to analyze a large number of mammogram scans efficiently. Models with shorter processing times can enable quicker decision-making and improve patient care. Inference time is influenced by factors such as model complexity, input data size, hardware resources, and optimization techniques. In breast cancer detection applications, a delay in obtaining the model's output could potentially lead to delayed diagnosis and treatment, underscoring the importance of

considering inference time alongside accuracy metrics like precision, recall, F1 score, and AUC when assessing the performance of hybrid deep learning and machine learning models for mammogram analysis (Hughes & Hughes, 2019).

This study evaluates the performance of the hybrid models by employing various aforementioned evaluation tools, namely accuracy, precision, recall, F1 score, AUC, and TIME. These metrics are used to measure the effectiveness and efficiency of the models in three distinct stages. In the first stage, referred to as training stage 1, the models are trained using the original mammogram images. This initial training phase allows the models to learn patterns and features present in the unaltered mammogram images. Moving on to the second stage, known as training stage 2, the models undergo further training using the same original mammogram images. However, this time, a set of image enhancement techniques is applied to the images before feeding them into the models. These enhancement techniques aim to improve the quality and clarity of the mammogram images, potentially leading to better model performance. Once the hybrid models have completed the training stages, they progress to the final stage, called the application stage. In this stage, the models are tested on a new set of images that were not used during the training process. The purpose of this stage is to assess the efficiency and generalization capabilities of the hybrid models on unseen data. By applying the models to this new image set, researchers can gauge their performance in real-world scenarios and evaluate their ability to accurately detect and classify relevant features in mammograms.

Chapter Four: Results

In the results section, this study systematically presents the findings across four distinct phases. The first phase focused on analyzing the CLAHE enhancement parameters, shedding light on the impact of CLAHE enhancement on the hybrid model's performance. This analysis contributed to a deeper understanding of image enhancement techniques in the context of breast cancer detection from mammogram images. Subsequently, the study evaluated the performance of the hybrid model using the original (unenhanced) data, establishing a foundational benchmark for further analyses. This assessment provided valuable insights into the model's performance without any image enhancement techniques applied. In the following phase, several image enhancement techniques, including morphological erosion, Gaussian Blur, CLAHE, and unsharp mask, were applied to the model. The aim was to investigate potential performance shifts and emphasize the significance of image clarity in improving diagnostic accuracy. Finally, the model was deployed in a real-world application, enabling the evaluation of its effectiveness within an actual clinical environment. This assessment bridged the gap between theoretical robustness and practical application, offering valuable insights into the model's real-world performance and its suitability for clinical use.

3.1 Contrast-Limited Adaptive Histogram Equalization (CLAHE)

The evaluation of the Peak Signal to Noise Ratio (PSNR) and the Effective Measure of Enhancement (EME) metrics for the CLAHE algorithm is presented in Table 9 and Figure 25. This evaluation was conducted using various ClipLimit values, specifically 2, 2.5, 3, 3.5, 4, 8, and 20, for three categories of mammogram images: Normal, Benign, and Malignant. Table 8 displays the PSNR values, which assess image quality by comparing the enhanced images with their original counterparts. Higher PSNR values indicate better image quality, indicating that the enhanced images closely resemble the originals. Across the "Normal," "Benign," and "Malignant" classifications, the PSNR values decrease as the ClipLimit values increase. This demonstrates that lower ClipLimit values (around 2) provide optimal enhancements for image quality. Conversely, higher ClipLimit values (around 20) lead to a reduction in image quality. Additionally, Table 8 presents the EME values, which measure the effectiveness of the enhancement algorithm in preserving edges and fine details in the enhanced images. Higher EME values indicate better edge preservation. For the "Normal" category, the EME values initially increase as the ClipLimit values increase up to around 4, indicating the best edge preservation. However, beyond ClipLimit 4, the

EME values start to decrease, suggesting that extreme ClipLimit values may result in a loss of edge information. Similar trends are observed in the "Benign" and "Malignant" categories, with increasing EME values up to ClipLimit 4 and slight decreases beyond that point.

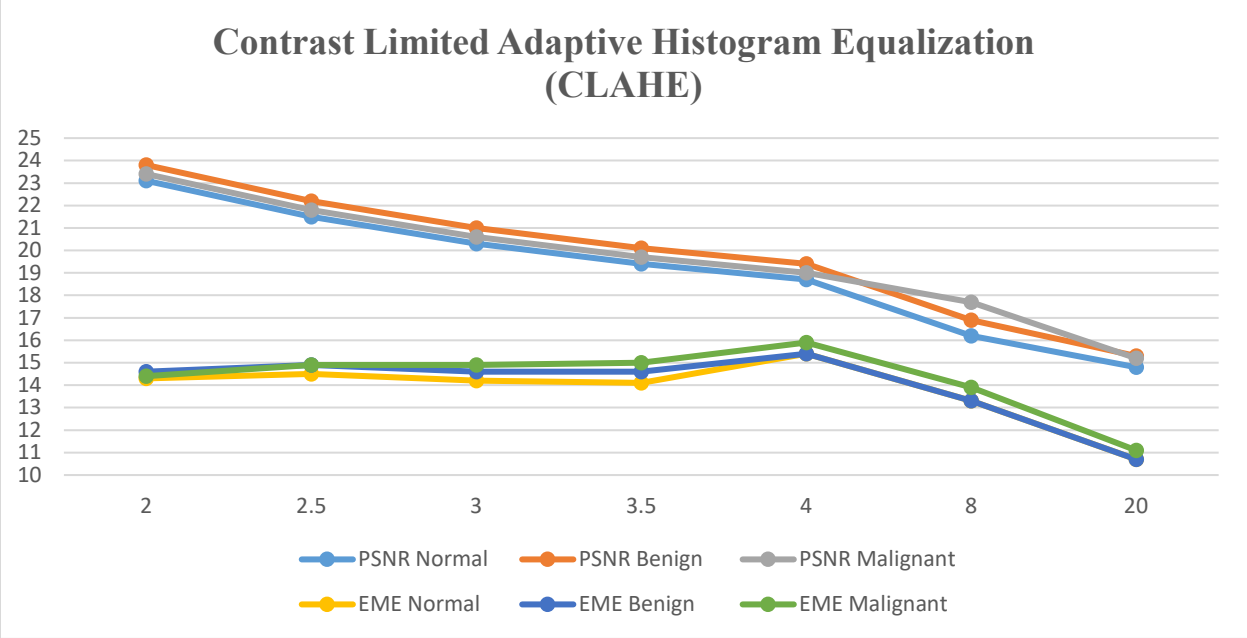

Figure 25: The interplay between Peak Signal-to-Noise Ratio (PSNR) and the Exponential Mean Error (EME) metrics, in conjunction with the ClipLimit parameter, to determine the optimal value for applying Contrast Limited Adaptive Histogram Equalization (CLAHE) in image enhancement.

Table 9: EME and PSNR values for data corresponding to different contrast thresholds.

Parameters	**Data Types**	**CLAHE – ClipLimit**						
		2	**2.5**	**3**	**3.5**	**4**	**8**	**20**
PSNR	**Normal**	23.1*	21.5	20.3	19.4	18.7	16.2	14.8
	Benign	23.8*	22.2	21.0	20.1	19.4	16.9	15.3
	Malignant	23.4*	21.8	20.6	19.7	19.0	17.7	15.2
EME	**Normal**	14.3	14.5	14.2	14.1	15.4*	13.3	10.7
	Benign	14.6	14.9	14.6	14.6	15.4*	13.3	10.7
	Malignant	14.4	14.9	14.9	15.0	15.9*	13.9	11.1

PSNR: Peak Signal to Noise Ratio; EME: Effective Measure of Enhancement.

*: Highest value in PSNR and EME.

3.1 Performance of Hybrid Artificial Intelligence (AI) Models without Enhancement techniques

Figure 26 presents a comprehensive evaluation of the performance of various machine learning classifiers in combination with two deep learning models, namely VGG16 and DenseNet121. These deep learning models were specifically employed for feature extraction from mammogram images. The extracted features were then subjected to classification using four different algorithms: random forest, gradient boosting, support vector machines, and logistic regression. The main objective of this analysis was to assess the effectiveness of each approach in classifying benign and malignant findings based on the original (unenhanced) data. To achieve this, multiple evaluation metrics were considered, including Accuracy, Precision, Recall, F1-Score, Area under the Curve (AUC), and the time taken for training and prediction.

The evaluation of various hybrid models for predicting benign breast cancer from original mammogram images revealed distinct performance characteristics. Among the models assessed, the Gradient Boosting (GB) Classifier with DenseNet121 Feature Extractor emerged as the top performer, achieving the highest accuracy of 0.869 and an impressive AUC of 0.858. These results demonstrate its exceptional discriminative power in distinguishing benign cases from other classes. However, its precision and recall ranked 4th and 3rd, respectively, indicating room for improvement in reducing false positives and false negatives. Consequently, its F1-score, which balances precision and recall, ranked 5th, indicating a trade-off between these metrics. Despite its strong predictive performance, this model exhibited a relatively long predictive time of 187.15 seconds, ranking 7th in terms of time efficiency.

Closely following the top performer was the Gradient Boosting (GB) Classifier with VGG16 Feature Extractor, achieving the second-highest accuracy of 0.865 and an AUC of 0.841. These results indicate its strong ability to differentiate benign cases. However, it's precision and recall ranked 6th and 5th, respectively, suggesting higher rates of misclassifications and missed true positive cases. As a result, its F1-score was the lowest at 0.432, highlighting an imbalance between precision and recall. Notably, this model exhibited the longest predictive time of 261.52 seconds, potentially limiting its practicality in time-sensitive scenarios. The Logistic Regression (LR) Classifier with DenseNet121 Feature Extractor attained the 3rd highest accuracy of 0.856 and the highest F1-score of 0.540, indicating a well-balanced trade-off between precision and recall. Its

recall ranked 2nd, suggesting effectiveness in detecting true positive benign cases, but its precision was the lowest, indicating a higher tendency to misclassify non-benign cases as benign. With an AUC of 0.837 and the 2nd shortest predictive time of 1.14 seconds, this model provided a relatively efficient option for practical applications.

Among the evaluated models, the Random Forest (RF) Classifier with DenseNet121 Feature Extractor achieved the 4th highest accuracy and F1-score. Notably, it demonstrated the highest precision of 0.656, indicating a good ability to minimize false positives. However, its recall ranked 4th, suggesting a higher rate of false negatives. With an AUC of 0.836 and a predictive time of 22.46 seconds, which ranked 4th in terms of time efficiency, this model offered moderate efficiency. The Logistic Regression (LR) Classifier with VGG16 Feature Extractor ranked 5th in accuracy but exhibited the highest recall, indicating effectiveness in detecting true positive benign cases. However, its precision was the lowest, suggesting a higher tendency to misclassify non-benign cases as benign. With the 2nd highest F1-score and the shortest predictive time of 0.80 seconds, this model provided a time-efficient option, although with a lower AUC of 0.814.

The Support Vector Machine (SVM) Classifier with DenseNet121 Feature Extractor ranked 6th in accuracy but achieved the 2nd highest precision, demonstrating its ability to minimize false positives. However, its recall ranked 6th, indicating a higher rate of false negatives. With an AUC of 0.848, which was the 2nd highest, and a predictive time of 14.97 seconds, ranking 3rd in time efficiency, this model offered a relatively efficient option with strong discriminative power. The Random Forest (RF) Classifier with VGG16 Feature Extractor obtained the 7th rank in accuracy and F1-score. It exhibited a moderate precision but the lowest recall, indicating a significant rate of missed true positive benign cases. Additionally, it achieved the lowest AUC of 0.806, suggesting a lower ability to distinguish benign cases from other classes. Its predictive time of 36.21 seconds ranked 5th in time efficiency. Lastly, the Support Vector Machine (SVM) Classifier with VGG16 Feature Extractor showed the lowest performance across multiple metrics, including accuracy, F1-score, and recall. Although it's precision ranked 3rd, its recall was the lowest, indicating a significant rate of missed true positive benign cases. With an AUC of 0.834, it demonstrated a relatively lower ability to distinguish benign cases, and its predictive time of 52.29 seconds ranked 6th in time efficiency.

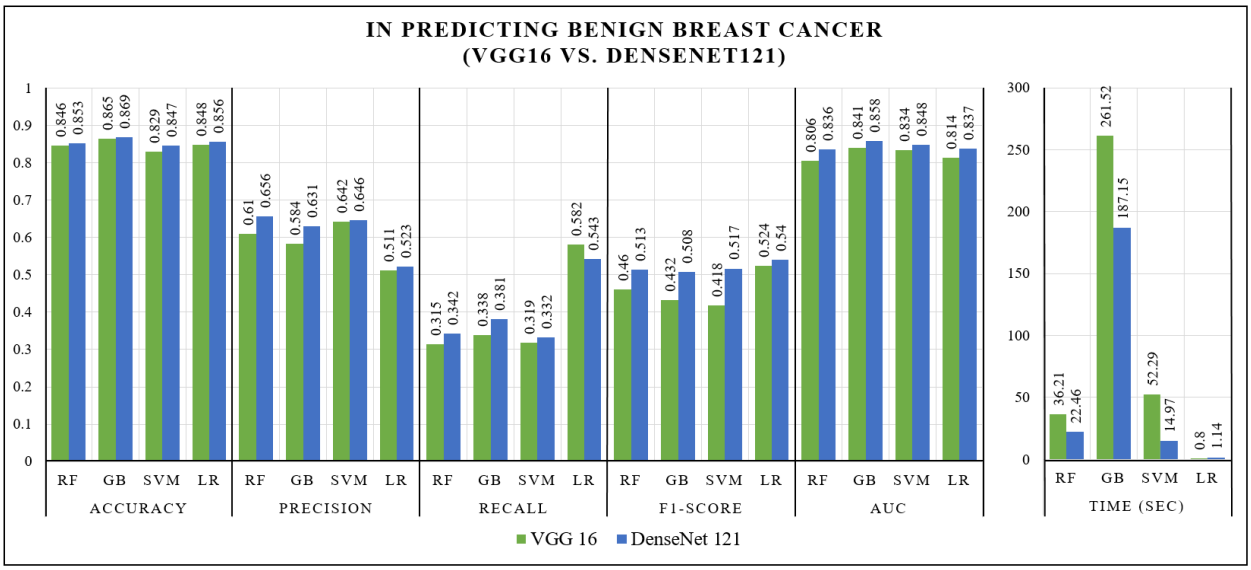

Figure 26: A comparative evaluation of the VGG16 and DenseNet121 feature extraction techniques, in conjunction with all the proposed classification models, for the task of predicting benign breast cancer cases.

When predicting cases of malignant breast cancer, as shown in Figure 27, the Logistic Regression (LR) Classifier with DenseNet121 Feature Extractor emerged as the top performer, achieving the highest accuracy of 0.872 and the highest recall of 0.564. This model demonstrated an exceptional ability to correctly identify true positive malignant cases, minimizing missed detections. However, it's precision of 0.579, ranking 6th, suggests a higher tendency to misclassify non-malignant cases as malignant, leading to a higher rate of false positives. Nonetheless, it achieved an F1-score of 0.560 and an AUC of 0.843, both ranking 2nd, indicating a well-balanced trade-off between precision and recall, as well as strong discriminative power. Notably, this model exhibited the highest time efficiency, with a predictive time of 0.1 seconds, making it highly practical for time-sensitive applications.

The Gradient Boosting (GB) Classifier with DenseNet121 Feature Extractor secured the second-highest accuracy of 0.871 but ranked lower in other metrics. Its precision of 0.625 and recall of 0.412, ranking 5th and 4th respectively, suggest a moderate tendency to misclassify both malignant and non-malignant cases. Consequently, its F1-score of 0.544 and AUC of 0.841, ranking 4th, reflect a less balanced trade-off between precision and recall, as well as lower discriminative power compared to the top performer. However, this model exhibited the longest predictive time of

122.78 seconds, potentially limiting its practicality in time-sensitive scenarios. The Random Forest (RF) Classifier with DenseNet121 Feature Extractor achieved the 3rd highest accuracy of 0.867 and the highest precision of 0.710, demonstrating its ability to effectively identify non-malignant cases and minimize false positives. However, its recall of 0.369, ranking 7th, indicates a significant rate of missed true positive malignant cases, resulting in a high rate of false negatives. It's F1-score of 0.553 and AUC of 0.842, both ranking 3rd, suggest a moderate trade-off between precision and recall, as well as decent discriminative power. With a predictive time of 15.58 seconds, ranking 6th, this model offered moderate time efficiency.

The Gradient Boosting (GB) Classifier with VGG16 Feature Extractor ranked 4th in accuracy at 0.856 but exhibited lower performance in other metrics. Its precision of 0.646 and recall of 0.398, ranking 4th and 5th respectively, suggest moderate misclassification rates for both malignant and non-malignant cases. Consequently, it's F1-score of 0.480, ranking 6th, reflects a poor trade-off between precision and recall. Additionally, it achieved the lowest AUC of 0.822, indicating the weakest discriminative power among the evaluated models. With a predictive time of 78.38 seconds, ranking 7th, this model was the second slowest, potentially limiting its practical utility. The Support Vector Machine (SVM) Classifier with DenseNet121 Feature Extractor achieved the 5th highest accuracy of 0.855 but demonstrated a well-balanced performance in other metrics. Its precision of 0.671 and recall of 0.486, ranking 2nd in both, suggest a moderate tendency to misclassify both malignant and non-malignant cases, but with a better balance compared to other models. Consequently, it achieved the highest F1-score of 0.583 and the highest AUC of 0.855, indicating the strongest overall trade-off between precision and recall, as well as the highest discriminative power. With a predictive time of 11.71 seconds, ranking 4th, this model offered reasonable time efficiency.

The Logistic Regression (LR) Classifier with VGG16 Feature Extractor ranked 6th in accuracy at 0.852 and exhibited lower performance in other metrics. Its precision of 0.560, ranking 7th, suggests a higher tendency to misclassify non-malignant cases as malignant, leading to a higher rate of false positives. Its recall of 0.460, ranking 3rd, indicates a moderate rate of missed true positive malignant cases. It's F1-score of 0.531 and AUC of 0.834, ranking 5th and 6th respectively, reflect a poor trade-off between precision and recall, as well as lower discriminative power. However, with a predictive time of 0.2 seconds, ranking 2nd, this model offered excellent

time efficiency. The Random Forest (RF) Classifier with VGG16 Feature Extractor ranked 7th in accuracy at 0.850 and exhibited the lowest performance in several metrics. Its precision of 0.652, ranking 3rd, suggests a moderate ability to identify non-malignant cases, minimizing false positives. However, its recall of 0.318 was the lowest among all models, indicating a significant rate of missed true positive malignant cases, resulting in a high rate of false negatives. Consequently, it's F1-score of 0.471 and AUC of 0.827, both ranking 7th, reflect the poorest trade-off between precision and recall and the weakest discriminative power. With a predictive time of 10.56 seconds, ranking 3rd, this model offered reasonable time efficiency. Finally, the Support Vector Machine (SVM) Classifier with VGG16 Feature Extractor exhibited the lowest accuracy of 0.846 and the lowest performance in several other metrics. Its precision of 0.552 was the lowest, suggesting a higher tendency to misclassify non-malignant cases as malignant, leading to a higher rate of false positives. Its recall of 0.392, ranking 6th, indicates a moderate rate of missed true positive malignant cases. Consequently, its F1-score of 0.465 was the lowest, reflecting the poorest trade-off between precision and recall. It's AUC of 0.836, ranking 5th, suggests moderate discriminative power. With a predictive time of 13.42 seconds, ranking 5th, this model offered moderate time efficiency.

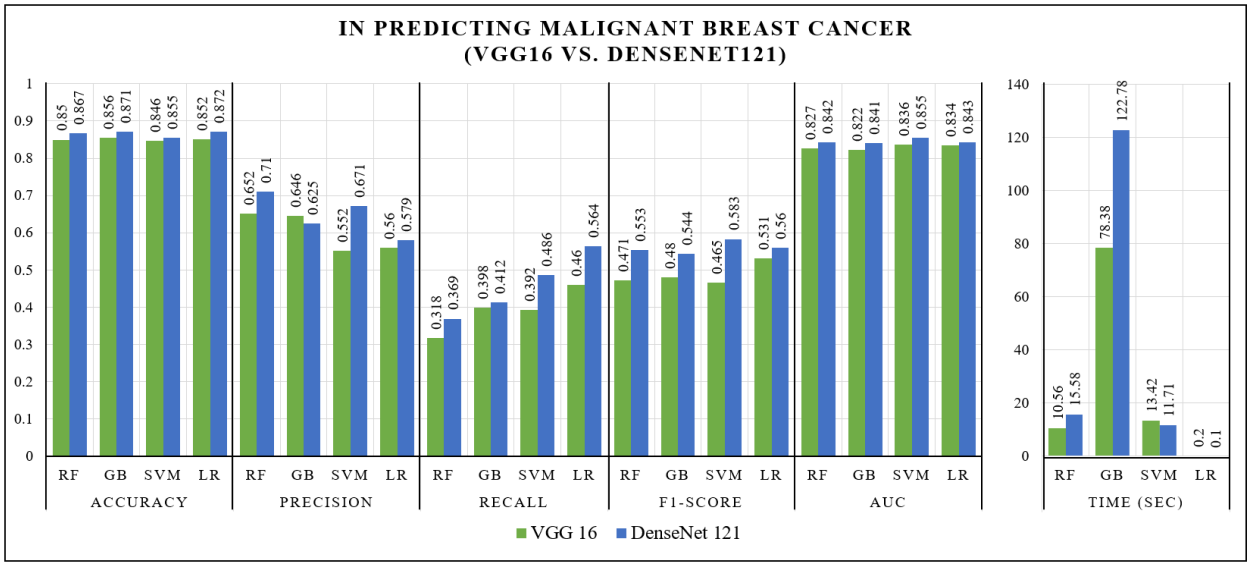

Figure 27: A comparative evaluation of the predictive performance exhibited by all the proposed classification models when employing the vgg16 feature extractor versus the densenet121 feature extractor for the task of identifying malignant breast cancer cases.

3.2 Performance of Hybrid Artificial Intelligence (AI) Models with Enhancement techniques

Figure 28 displays the performance evaluation of the previously mentioned hybrid AI models on enhanced data. The data underwent pre-processing techniques including morphological erosion, Gaussian Blur, CLAHE, and unsharp mask to predict cases as benign or malignant. Specifically, CLAHE was applied with a clip limit of 4. The evaluation metrics consisted of Accuracy, Precision, Recall, F1-Score, AUC, and the time taken for training and prediction (Time/S). In predicting benign breast cancer cases from enhanced mammogram images, the Logistic Regression (LR) Classifier with DenseNet121 Feature Extractor emerged as the top performer. It achieved the highest accuracy of 0.991, precision of 0.996, F1-score of 0.989, and AUC of 0.999. These exceptional metrics indicate its remarkable ability to correctly identify benign cases while minimizing misclassifications of both benign and non-benign cases. Furthermore, its recall of 0.978, ranking 2nd, suggests a minimal rate of missed true positive benign cases. Notably, this model exhibited the 2nd shortest predictive time of 0.15 seconds, showcasing high efficiency for practical applications. Following closely behind the top performer was the Support Vector Machine (SVM) Classifier with DenseNet121 Feature Extractor, achieving the 2nd highest accuracy of 0.986, precision of 0.991, and F1-score of 0.987. It shared the highest AUC of 0.999 with the top model, indicating exceptional discriminative power in distinguishing benign cases from other classes. However, its recall of 0.946, ranking 4th, suggests a slightly higher rate of missed true positive benign cases compared to the top model. With a predictive time of 8.08 seconds, ranking 5th, this model offered moderate time efficiency.

The Logistic Regression (LR) Classifier with VGG16 Feature Extractor attained the 3rd highest accuracy of 0.985 and the highest recall of 0.979, demonstrating its effectiveness in detecting true positive benign cases and minimizing false negatives. Its precision of 0.986 and F1-score of 0.987, both ranking 3rd, indicate a well-balanced trade-off between precision and recall. However, it's AUC of 0.987, ranking 4th, suggests slightly lower discriminative power compared to the top performers. Notably, this model exhibited the shortest predictive time of 0.13 seconds, making it the most time-efficient option among the evaluated models. The Support Vector Machine (SVM) Classifier with VGG16 Feature Extractor ranked 4th in accuracy at 0.975 and demonstrated lower performance in other metrics. Its precision of 0.981 and recall of 0.959, ranking 5th and 3rd

respectively, suggest a moderate tendency to misclassify both benign and non-benign cases. Consequently, it's F1-score of 0.841, ranking 5th, reflects a poor trade-off between precision and recall. Additionally, it's AUC of 0.986, ranking 5th, indicates lower discriminative power compared to the top performers. With a predictive time of 4.45 seconds, ranking 3rd, this model offered good time efficiency.

The Gradient Boosting (GB) Classifier with VGG16 Feature Extractor achieved the 5th highest accuracy of 0.973 but demonstrated lower performance in other metrics. Its precision of 0.978 and recall of 0.867, ranking 6th and 7th respectively, suggest a higher tendency to misclassify both benign and non-benign cases. Consequently, it's F1-score of 0.924, ranking 4th, reflects a poor trade-off between precision and recall. However, it's AUC of 0.987, ranking 4th, indicates decent discriminative power. With a predictive time of 58.25 seconds, ranking 7th, this model was the second slowest, potentially limiting its practical utility. The Random Forest (RF) Classifier with DenseNet121 Feature Extractor achieved the 6th highest accuracy of 0.969 but exhibited lower performance in other metrics. Its precision of 0.902, ranking 7th, suggests a higher tendency to misclassify non-benign cases as benign, leading to a higher rate of false positives. Its recall of 0.881, ranking 5th, indicates a moderate rate of missed true positive benign cases. Consequently, it's F1-score of 0.836, ranking 6th, reflects a poor trade-off between precision and recall. However, it's AUC of 0.994, ranking 2nd, indicates strong discriminative power. With a predictive time of 15.43 seconds, ranking 6th, this model offered moderate time efficiency.

The Gradient Boosting (GB) Classifier with DenseNet121 Feature Extractor achieved the 7th highest accuracy of 0.957 but exhibited lower performance in other metrics. Its precision of 0.985 and recall of 0.872, ranking 4th and 6th respectively, suggest a moderate tendency to misclassify both benign and non-benign cases. Consequently, it's F1-score of 0.939, ranking 3rd, reflects a moderate trade-off between precision and recall. It's AUC of 0.992, ranking 3rd, indicates strong discriminative power. However, this model had the longest predictive time of 141.05 seconds, potentially limiting its practical applicability in time-sensitive scenarios. Lastly, the Random Forest (RF) Classifier with VGG16 Feature Extractor exhibited the lowest performance across multiple metrics. With an accuracy of 0.952, precision of 0.985, and recall of 0.803, all ranking lowest, this model demonstrated a higher tendency to misclassify both benign and non-benign cases, leading to higher rates of false positives and false negatives. Consequently, it's F1-score of

0.814 and AUC of 0.976, both ranking lowest, reflect the poorest trade-off between precision and recall, and the weakest discriminative power. With a predictive time of 4.83 seconds, ranking 4th, this model offered moderate time efficiency.

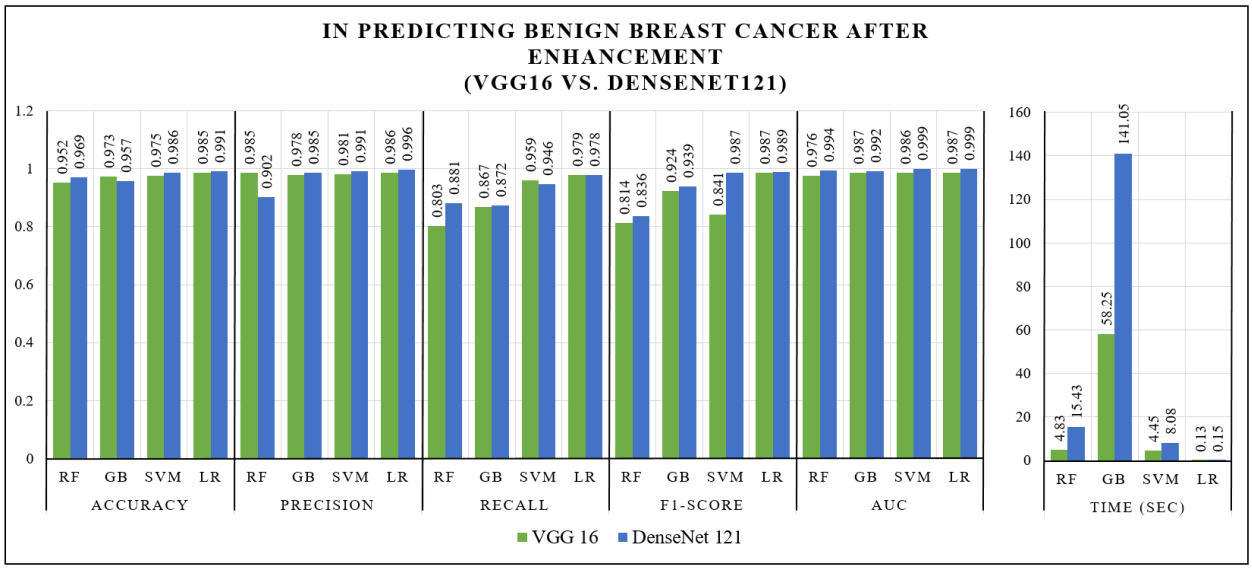

Figure 28: A comparative evaluation of the predictive performance achieved by all the proposed classification models when identifying benign breast cancer cases, contrasting the results

In the task of predicting malignant breast cancer cases from enhanced mammogram images, as shown in Figure 29, the Logistic Regression (LR) Classifier with DenseNet121 Feature Extractor emerged as the top performer across all metrics. It achieved exceptional results with the highest accuracy, precision, recall, F1-score, and AUC, all measuring at 0.995 and above. These outstanding metrics indicate its unparalleled ability to correctly identify malignant cases while minimizing misclassifications of both malignant and non-malignant cases. Furthermore, this model demonstrated a highly efficient predictive time of only 0.22 seconds, ranking second in terms of speed, making it well-suited for practical applications. The Support Vector Machine (SVM) Classifier with DenseNet121 Feature Extractor secured the second-highest performance across multiple metrics, including accuracy (0.992), precision (0.992), F1-score (0.982), and AUC (0.998). Its recall of 0.966, ranking second, suggests a minimal rate of missed true positive malignant cases. Although its predictive time of 19.03 seconds ranked fifth in speed, this model still offered moderate time efficiency while maintaining strong predictive performance and discriminative power.

The Logistic Regression (LR) Classifier with VGG16 Feature Extractor achieved the third-highest accuracy of 0.983 and an F1-score of 0.955. However, its precision of 0.960, ranking seventh, suggests a higher tendency to misclassify non-malignant cases as malignant, leading to a higher rate of false positives. Its recall of 0.950, ranking fourth, indicates a moderate rate of missed true positive malignant cases. Nevertheless, its AUC of 0.996, ranking third, demonstrates strong discriminative power. Notably, this model exhibited the shortest predictive time of 0.08 seconds, making it the most time-efficient option among the evaluated models. The Gradient Boosting (GB) Classifier with DenseNet121 Feature Extractor achieved the 4th highest accuracy of 0.982 but displayed lower performance in other metrics. Its precision of 0.979 and recall of 0.892, ranking 5th and 6th respectively, indicate a moderate tendency to misclassify both malignant and non-malignant cases. Consequently, it's F1-score of 0.947, ranking 5th, reflects a poor trade-off between precision and recall. However, it's AUC of 0.996, ranking 3rd, demonstrates strong discriminative power. This model exhibited the longest predictive time of 219.23 seconds, potentially limiting its practical utility in time-sensitive scenarios.

The Support Vector Machine (SVM) Classifier with VGG16 Feature Extractor achieved the 5th highest accuracy of 0.980 but demonstrated mixed performance in other metrics. Its precision of 0.989, ranking 3rd, suggests a low tendency to misclassify non-malignant cases as malignant, minimizing false positives. Its recall of 0.961, ranking 3rd, indicates a minimal rate of missed true positive malignant cases. However, its F1-score of 0.848 was the lowest, reflecting a poor trade-off between precision and recall. Additionally, it's AUC of 0.994, ranking 4th, indicates strong discriminative power. With a predictive time of 3.74 seconds, ranking 3rd, this model offered good time efficiency. The Gradient Boosting (GB) Classifier with VGG16 Feature Extractor achieved the 6th highest accuracy of 0.976 but demonstrated lower performance in other metrics. Its precision of 0.977 and recall of 0.874, ranking 6th and 7th respectively, suggest a higher tendency to misclassify both malignant and non-malignant cases. Consequently, it's F1-score of 0.933, ranking 6th, reflects a poor trade-off between precision and recall. Additionally, it's AUC of 0.991, ranking 6th, indicates lower discriminative power compared to the top performers. With a predictive time of 43.94 seconds, ranking 7th, this model was the second slowest, potentially limiting its practical applicability.

The Random Forest (RF) Classifier with DenseNet121 Feature Extractor achieved the 7th highest accuracy of 0.973 but displayed mixed performance in other metrics. Its precision of 0.988, ranking 4th, suggests a low tendency to misclassify non-malignant cases as malignant, minimizing false positives. However, its recall of 0.895, ranking 5th, indicates a moderate rate of missed true positive malignant cases. The F1-score of 0.950, ranking 4th, reflects a moderate trade-off between precision and recall. Additionally, it's AUC of 0.996, ranking 3rd, demonstrates strong discriminative power. With a predictive time of 25.49 seconds, ranking 6th, this model offered moderate time efficiency. On the other hand, the Random Forest (RF) Classifier with VGG16 Feature Extractor exhibited the lowest performance across multiple metrics. With an accuracy of 0.959, precision of 0.988, and recall of 0.825, all ranking lowest, this model demonstrated a higher tendency to misclassify both malignant and non-malignant cases, resulting in higher rates of false positives and false negatives. Consequently, its F1-score of 0.893 was the second lowest, reflecting a poor trade-off between precision and recall. It's AUC of 0.992, ranking 5th, indicates lower discriminative power compared to the top performers. With a predictive time of 5.33 seconds, ranking 4th, this model offered moderate time efficiency.

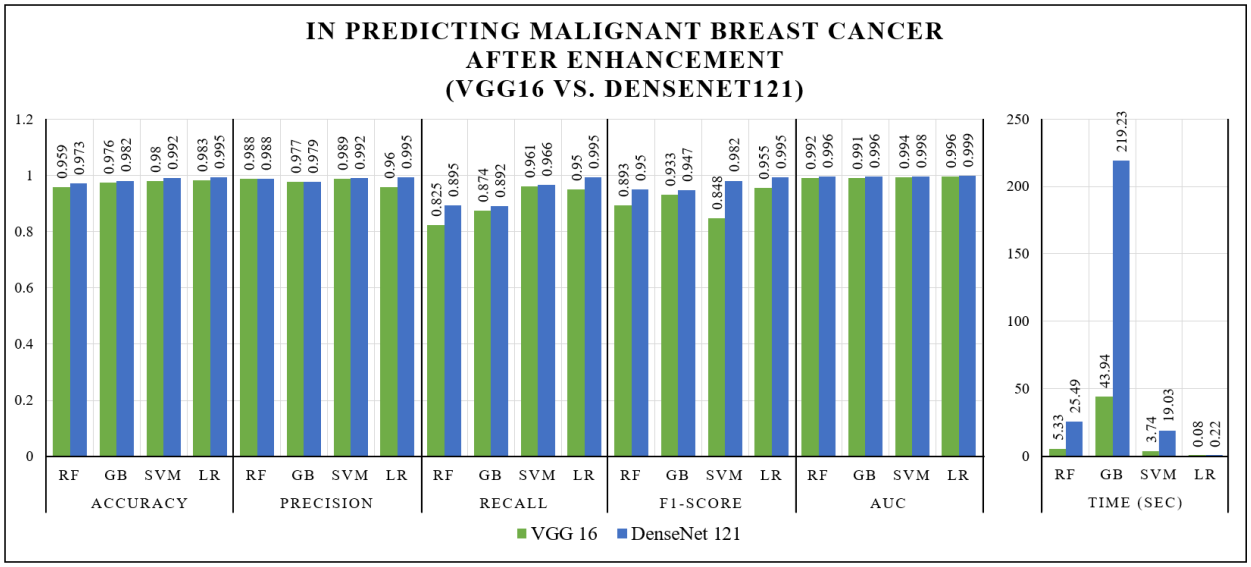

Figure 29: A comparative evaluation of the predictive performance achieved by all classification models when identifying malignant breast cancer cases, contrasting the results obtained using the VGG16 feature extractor against those obtained with the DenseNet121 feature extractor, after applying image enhancement techniques to the input data.

3.3 Validation of Hybrid Artificial Intelligence (AI) Models with Enhancement techniques

Following the methodology, a subset of 10% of the mammogram images was reserved for verifying the proposed models' applicability in image diagnosis. This subset consisted of 400 cases of benign breast cancer and 400 cases of malignant breast cancer, confirmed by biopsy reports. However, during the application of the proposed models to this dataset, the images were not classified and instead placed together in a single folder, regardless of their benign or malignant nature. Subsequently, the proposed models, along with the mentioned preprocessing enhancements, were applied to these images, and their predictive capabilities were evaluated using performance metrics. In Figure 30, the Logistic Regression (LR) Classifier with DenseNet121 Feature Extractor emerged as the top performer, achieving the highest accuracy of 0.967, prediction percentage of 95.2%, recall of 0.975, F1-score of 0.939, and AUC of 0.999. These exceptional metrics demonstrate its ability to accurately identify benign cases while minimizing false negatives. However, it's precision of 0.886, ranking 5th, suggests a moderate tendency to misclassify non-benign cases as benign, leading to a higher rate of false positives. Importantly, this model exhibited the second shortest predictive time of 0.06 seconds, making it highly efficient for practical use.

The Random Forest (RF) Classifier with DenseNet121 Feature Extractor achieved the second highest accuracy of 0.950 and shared the highest prediction percentage of 95.2%. Its precision of 0.933, ranking 4th, indicates a moderate tendency to misclassify non-benign cases as benign. However, its recall of 0.824, ranking 4th, suggests a higher rate of missed true positive benign cases. The F1-score of 0.903, ranking 3rd, reflects a moderate trade-off between precision and recall. Additionally, it's AUC of 0.984, ranking 4th, demonstrates strong discriminative power. With a predictive time of 6.17 seconds, ranking 6th, this model offered moderate time efficiency. The Logistic Regression (LR) Classifier with VGG16 Feature Extractor achieved the third highest accuracy of 0.942 and the second highest prediction percentage of 94.0%. Its precision of 0.872, ranking 6th, suggests a higher tendency to misclassify non-benign cases as benign. However, its recall of 0.944, ranking 2nd, indicates a minimal rate of missed true positive benign cases. The F1-score of 0.907, ranking 2nd, reflects a good trade-off between precision and recall. Additionally,

it's AUC of 0.986, ranking 3rd, demonstrates strong discriminative power. Notably, this model exhibited the shortest predictive time of 0.05 seconds, making it the most time-efficient option.

The Gradient Boosting (GB) Classifier with DenseNet121 Feature Extractor and the Support Vector Machine (SVM) Classifier with DenseNet121 Feature Extractor achieved an accuracy of 0.933, ranking 4th. The GB Classifier demonstrated a higher precision of 0.964, ranking 2nd, but a lower recall of 0.794, ranking 5th. In contrast, the SVM Classifier had a lower precision of 0.814, ranking 7th, but a higher recall of 0.765, ranking 7th. The GB Classifier's F1-score of 0.871, ranking 4th, indicated a better trade-off between precision and recall compared to the SVM Classifier's F1-score of 0.866, ranking 5th. The GB Classifier had a lower AUC of 0.973, ranking 6th, while the SVM Classifier achieved an AUC of 0.980, ranking 5th. However, the GB Classifier had a significantly longer predictive time of 56.42 seconds as the longest, whereas the SVM Classifier had a moderate predictive time of 3.23 seconds, ranking 5th.

The Gradient Boosting (GB) Classifier with VGG16 Feature Extractor achieved an accuracy of 0.925, ranking 5th, and exhibited a high precision of 0.965, ranking as the highest. However, its recall of 0.777, ranking 6th, suggested a higher rate of missed true positive benign cases. Consequently, it's F1-score of 0.862, ranking 6th, indicated a poor trade-off between precision and recall. Additionally, it's AUC of 0.947, ranking 7th, indicated lower discriminative power compared to the top performers. With a predictive time of 23.30 seconds, ranking 7th, this model was the second slowest, potentially limiting its practical applicability. The Support Vector Machine (SVM) Classifier with VGG16 Feature Extractor achieved an accuracy of 0.916, ranking 6th, and a prediction percentage of 91.7%, ranking 4th. While its recall of 0.893, ranking 3rd, suggested a moderate rate of missed true positive benign cases, its precision of 0.722, ranking as the lowest, indicated a higher tendency to misclassify non-benign cases as benign, leading to a higher rate of false positives. Consequently, its F1-score of 0.838 was the lowest, reflecting a poor trade-off between precision and recall. However, it's AUC of 0.988, ranking 2nd, demonstrated strong discriminative power. With a predictive time of 1.78 seconds, ranking 3rd, this model offered good time efficiency.

Finally, the Random Forest (RF) Classifier with VGG16 Feature Extractor demonstrated the lowest performance across multiple metrics. With an accuracy of 0.833, prediction percentage of 90.4%, precision of 0.958, recall of 0.638, and F1-score of 0.766, all ranking as the lowest, this

model exhibited a higher tendency to misclassify both benign and non-benign cases, resulting in higher rates of false positives and false negatives. However, it's AUC of 0.930, ranking 4th, indicated decent discriminative power. With a predictive time of 1.93 seconds, ranking 5th, this model offered moderate time efficiency.

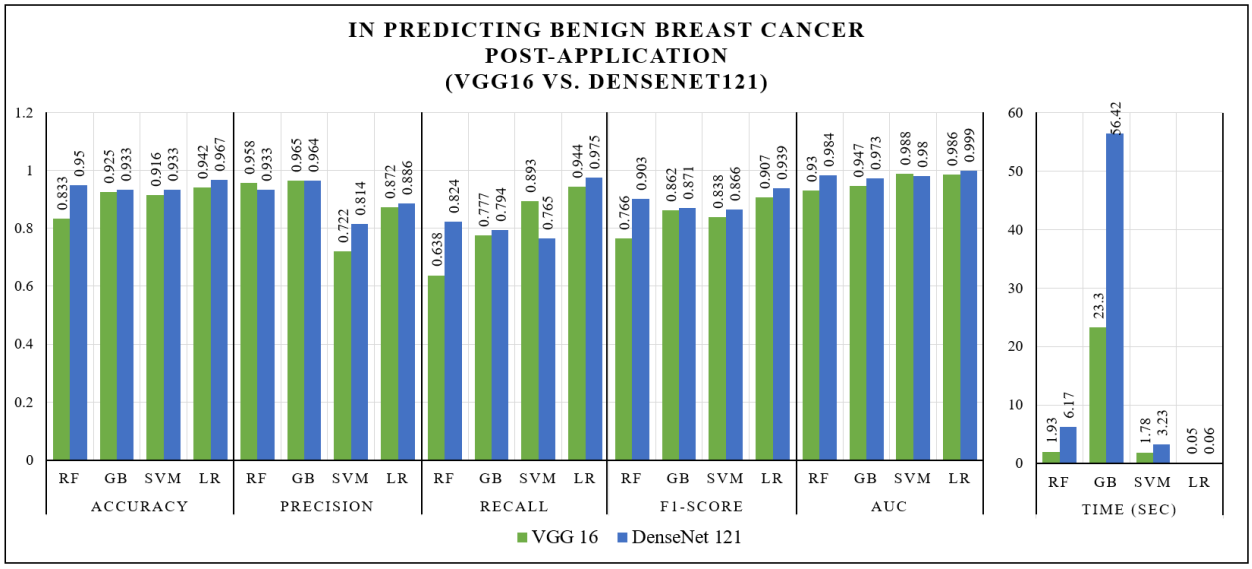

Figure 30: A comparative analysis of the predictive performance achieved by all classification models when identifying benign breast cancer cases from new, unseen mammogram images, contrasting the results obtained using the VGG16 feature extractor against those obtained with the DenseNet121 feature extractor.

When it comes to predicting malignant breast cancer cases, as shown in Figure 31, the Logistic Regression (LR) Classifier with DenseNet121 Feature Extractor emerged as the top performer, achieving the highest accuracy of 0.955, prediction percentage of 95.8%, recall of 0.946, F1-score of 0.933, and AUC of 0.998. These exceptional metrics indicate its remarkable ability to accurately identify malignant cases while minimizing false negatives. However, it's precision of 0.921, ranking 3rd, suggests a moderate tendency to misclassify non-malignant cases as malignant, resulting in a higher rate of false positives. Importantly, this model exhibited the 2nd shortest predictive time of 0.17 seconds, making it highly efficient for practical applications. The Logistic Regression (LR) Classifier with VGG16 Feature Extractor secured the second highest accuracy of 0.937 and prediction percentage of 94.4%. Its precision of 0.902, ranking 5th, indicates a higher tendency to misclassify non-malignant cases as malignant. However, its recall of 0.925, ranking

2nd, suggests a minimal rate of missed true positive malignant cases. The F1-score of 0.914, ranking 2nd, reflects a good trade-off between precision and recall. Additionally, it's AUC of 0.994, ranking 2nd, demonstrates strong discriminative power. Notably, this model exhibited the shortest predictive time of 0.06 seconds, making it the most time-efficient option.

Both the Gradient Boosting (GB) Classifier with DenseNet121 Feature Extractor and the Support Vector Machine (SVM) Classifier with DenseNet121 Feature Extractor achieved an accuracy of 0.928, ranking 3rd. The GB Classifier demonstrated a higher precision of 0.955, ranking 2nd, but a lower recall of 0.789, ranking 6th. On the other hand, the SVM Classifier had a lower precision of 0.916, ranking 4th, but a higher recall of 0.790, ranking 5th. As a result, the GB Classifier's F1-score of 0.862, ranking 3rd, indicates a better trade-off between precision and recall compared to the SVM Classifier's F1-score of 0.857, ranking 4th. However, the GB Classifier exhibited a lower AUC of 0.976, ranking 6th, while the SVM Classifier achieved an AUC of 0.990, ranking 4th. Furthermore, the GB Classifier had a significantly longer predictive time of 168.64 seconds, ranking as the longest, while the SVM Classifier had a moderate predictive time of 14.64 seconds, and ranking 5th.

The Random Forest (RF) Classifier with VGG16 Feature Extractor achieved an accuracy of 0.923, ranking 4th, and a prediction percentage of 93.1%, ranking 3rd. However, it's precision of 0.816, ranking 7th, indicates a higher tendency to misclassify non-malignant cases as malignant. The recall of 0.800, ranking 4th, suggests a moderate rate of missed true positive malignant cases. As a result, its F1-score of 0.811, ranking as the lowest, reflects a poor trade-off between precision and recall. Additionally, it demonstrated the lowest AUC of 0.948, indicating lower discriminative power compared to the top performers. On the positive side, with a predictive time of 4.10 seconds, ranking 4th, this model offered good time efficiency. The Gradient Boosting (GB) Classifier with VGG16 Feature Extractor achieved an accuracy of 0.921, ranking 5th, and a prediction percentage of 91.7%, ranking 4th. Although it exhibited the highest precision of 0.968, its recall of 0.750, ranking 7th, suggests a higher rate of missed true positive malignant cases. Consequently, it's F1-score of 0.845, ranking 5th, reflects a poor trade-off between precision and recall. Moreover, it's AUC of 0.974, ranking 7th, indicates lower discriminative power compared to the top performers. With a predictive time of 33.80 seconds, ranking 7th, this model was the second slowest, potentially limiting its practical applicability.

The Random Forest (RF) Classifier with DenseNet121 Feature Extractor achieved an accuracy of 0.919, ranking 6th, and a prediction percentage of 93.1%, ranking 3rd. Its precision of 0.894, ranking 6th, suggests a higher tendency to misclassify non-malignant cases as malignant. Additionally, its recall of 0.729, ranking as the lowest, indicates a higher rate of missed true positive malignant cases. Consequently, it's F1-score of 0.836, ranking 6th, reflects a poor trade-off between precision and recall. However, it demonstrated a good AUC of 0.982, ranking 5th, indicating good discriminative power. With a predictive time of 19.61 seconds, ranking 6th, this model offered moderate time efficiency. Finally, the Support Vector Machine (SVM) Classifier with VGG16 Feature Extractor exhibited the lowest accuracy of 0.901 and prediction percentage of 91.7%, ranking 4th. While its recall of 0.852, ranking 3rd, suggests a moderate rate of missed true positive malignant cases, its precision of 0.706, ranking as the lowest, indicates a higher tendency to misclassify non-malignant cases as malignant, leading to a higher rate of false positives. Consequently, its F1-score of 0.810 was the lowest, reflecting a poor trade-off between precision and recall. However, it's AUC of 0.991, ranking 3rd, demonstrates strong discriminative power. With a predictive time of 2.88 seconds, ranking 3rd, this model offered good time efficiency.

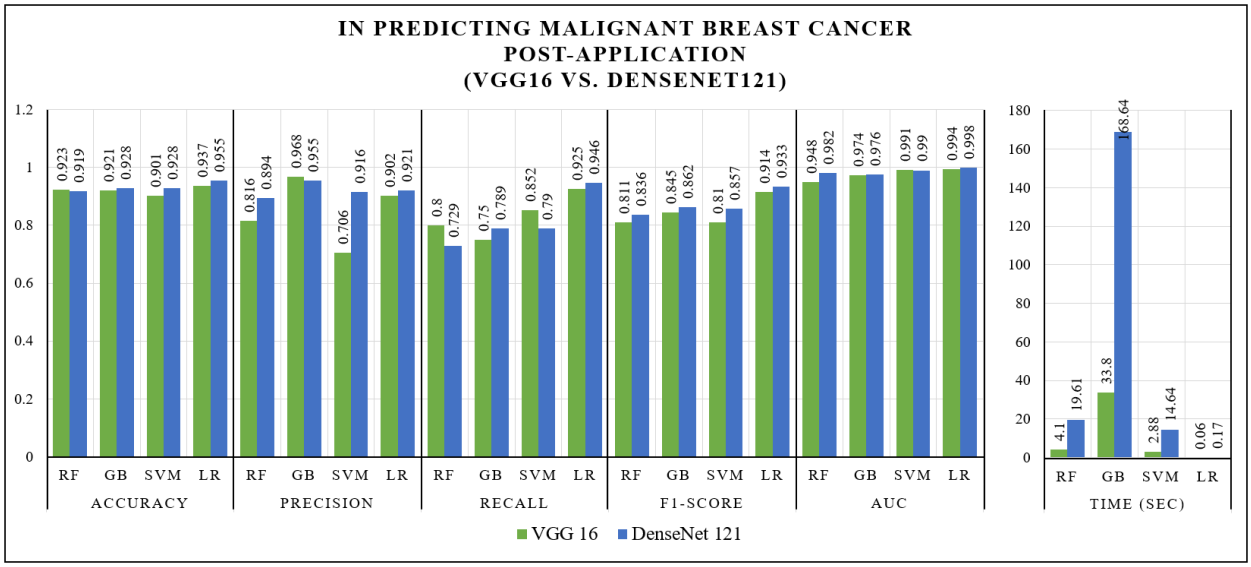

Figure 31: A comparative analysis of the predictive performance achieved by all classification models when identifying malignant breast cancer cases from new, unseen mammogram images, contrasting the results obtained using the VGG16 feature extractor against those obtained with the DenseNet121 feature extractor.

Chapter Five: Discussion

Within this chapter, the results will be examined and presented through comparisons, which are organized into three primary categories. The initial category focuses on analyzing the PSNR and EME outcomes, aiming to identify the optimal ClipLimit for implementing the CLAHE mammogram image enhancement technique. The second category delves into the performance evaluation of the proposed hybrid artificial intelligence models for predicting benign and malignant breast cancer. This evaluation encompasses various metrics, including Accuracy, Precision, Recall, F1-Score, AUC, and Time. Lastly, the final category explores the feasibility of applying these models in the field of diagnostic medicine and gauges their efficacy within real-world scenarios, thereby validating their potential as effective diagnostic tools.

4.1 Analysis of Contrast-Limited Adaptive Histogram Equalization (CLAHE)

The relationship between the clip limit and the corresponding values of PSNR (Peak Signal-to-Noise Ratio) and EME (Entropy-based Metric) in determining the optimal ClipLimit for utilizing Contrast Limited Adaptive Histogram Equalization (CLAHE). Notably, a ClipLimit of 2 yields the highest PSNR, which is significant as it measures the quality of the image reconstruction following contrast enhancement. This outcome can be attributed to the lower clip limit allowing for a more aggressive form of contrast limiting. Consequently, the algorithm adjusts the image's contrast more rigorously, leading to better preservation of fine image details (Pisano et al., 1998). The preservation of these details contributes to a higher PSNR, indicating superior image quality after the contrast enhancement process. This understanding is crucial in determining the ideal clip limit that balances the preservation of image details with visual quality enhancement (ABBOODI, 2014).

Conversely, adjusting the ClipLimit to 4 results in the highest EME, which measures the amount of information present in an image. A higher EME value signifies more diverse and complex information, often associated with improved visual quality. This outcome is due to the slightly higher clip limit (4) providing greater flexibility in contrast enhancement. The algorithm can adjust the image's contrast with more nuance and precision, leading to refined enhancement. Consequently, the overall visual quality is improved, resulting in a higher EME value. This

understanding is vital in determining the optimal ClipLimit that not only preserves image details but also contributes to an overall enhancement in visual quality, as indicated by the higher EME value (Intriago-Pazmiño et al., 2023).

In this study, the selection of ClipLimit = 4 was based on its ability to achieve a harmonious balance between two crucial aspects: preserving image details (high PSNR) and enhancing overall visual quality (high EME). This equilibrium ensures that the image processing methodology is optimized to prioritize both the preservation of intricate image details and the improvement of overall visual quality. By choosing ClipLimit = 4, the goal is to strike a balance where the contrast enhancement process retains fine image details (high PSNR) while contributing to an overall enhancement in visual quality (high EME). This balance is essential to ensure that the resulting images exhibit enhanced visual quality and preserved intricate details, both of which are critical aspects of effective image processing.

4.2 Evaluation of hybrid artificial intelligence (AI) models

In this particular section of the discussion, we will explore the analysis of performance results for a set of proposed hybrid artificial intelligence models designed to predict benign and malignant breast cancer. The evaluation process will consist of two stages: the training stage and the application stage. The training stage can be further divided into two sections. The first section focuses on training the hybrid models using the original mammogram database without employing any image enhancement techniques. The primary objective here is to determine which hybrid model performs the best in predicting benign and malignant breast cancer. The original mammogram database serves as the foundation for training these models, and a comprehensive assessment is conducted to identify the most effective model for the task. The second section of the training stage involves training the hybrid models using an enhanced mammogram database. This enhanced database is derived from the original mammogram dataset after applying various image enhancement techniques. The purpose of this section is to evaluate the impact of these image enhancement techniques on enhancing the performance of the proposed hybrid models. By training the models on the enhanced database, the analysis aims to ascertain how these techniques contribute to the accuracy and effectiveness of the models in predicting benign and malignant breast cancer. A comparative analysis is conducted among the hybrid models to identify the top-performing model in this context.

Following the training stage, the next step is the application stage. In this stage, the best-performing hybrid model is applied to a new, unclassified mammogram database. The objective is to validate the efficiency and effectiveness of the selected model on previously unseen data. By evaluating the model's performance on this new database, its predictive capabilities can be assessed in a real-world scenario. Furthermore, the results obtained from the best-performing hybrid model are compared with other hybrid models used in previous studies. This comparative analysis provides insights into the advancements and improvements achieved by the proposed hybrid models compared to existing approaches. It helps us understand the progress made by these models and their potential contributions to the field.

4.2.1 Training Stage 1: From Original Mammogram Images

Based on the previous findings, it is observed that all the proposed hybrid models demonstrate satisfactory performance in predicting benign breast cancer, with accuracy ranging from 0.829 to 0.869. Similarly, in predicting malignant breast cancer, the models exhibit accuracy ranging from 0.846 to 0.872 when utilizing the original mammogram images. The success of these models can be attributed to the approach employed in this study, which involved experimenting with a selection of hybrid models and identifying the best-performing one. However, slight variations in accuracy are noticeable among the selected hybrid models, primarily due to the classifier's characteristics and the chosen feature extractor. These differences arise from the specific algorithms employed by each model's classifier and the unique processing capabilities of the feature extractor when handling mammogram image data. Classifiers and feature extractors possess distinct strengths, weaknesses, and biases due to their underlying mathematical formulations and training methodologies.

4.2.1.1 Hybrid AI models based on VGG16 feature extractor in predicting benign breast cancer

The performance of the proposed classifiers utilizing the VGG16 feature extractor in predicting benign breast cancer from original mammogram images. Among the VGG16 models, the Gradient Boosting (GB) classifier emerges as the top performer for this task. It achieves a high accuracy of 0.865 and an AUC of 0.841, indicating its exceptional ability to accurately classify benign cases while effectively distinguishing them from non-benign cases. The success of the GB classifier can be attributed to the ensemble nature of gradient boosting, which combines multiple weak learners

to create a robust and accurate model. These findings align with the study conducted by Yao et al. (2019), which demonstrated the effectiveness of gradient boosting classifiers in breast cancer diagnosis using mammogram images, achieving an AUC of 0.91. However, it is important to note that the GB classifier exhibits a relatively low precision of 0.584 and recall of 0.338. This suggests a higher tendency to misclassify non-benign cases as benign (false positives) or benign cases as non-benign (false negatives), respectively. This could be attributed to the model's complexity and the potential for overfitting to the training data, as highlighted in the study by Ridgeway (1999) that emphasized the risk of overfitting in gradient boosting models. Additionally, the GB classifier has the longest predictive time of 261.52 seconds among all the models. This longer prediction time is due to the iterative nature of gradient boosting, making it computationally expensive, particularly for large datasets or real-time applications. This limitation was also observed in the study by Friedman (2001).

Contrasting the Logistic Regression (LR) classifier reveals a distinct set of strengths and weaknesses. It achieves the highest recall of 0.582 among all models and the highest F1-score of 0.524 among the VGG16 models. These metrics indicate its proficiency in accurately identifying positive cases (benign breast cancer) while maintaining a balanced trade-off between precision and recall. The LR classifier's performance can be attributed to the simplicity and interpretability of logistic regression models, which effectively capture underlying patterns in the data. These findings are supported by the study conducted by Brem et al. (2017), which demonstrated the effectiveness of logistic regression in breast cancer diagnosis using mammogram features, achieving an AUC of 0.84. Notably, the LR classifier stands out with the shortest predictive time of 0.80 seconds, making it highly computationally efficient for rapid decision-making. The computational efficiency of logistic regression models is highlighted in the study by Le Cessie and Van Houwelingen (1992). However, its lowest precision of 0.511 among all models indicates a higher likelihood of producing false positives, where non-benign cases may be misclassified as benign. This limitation may stem from the linear nature of logistic regression, which can struggle to capture complex relationships in the data, as discussed in the study by Dreiseitl and Ohno-Machado (2002).

In contrast, the Random Forest (RF) classifier stands out with the highest precision of 0.610 among the VGG16 models, demonstrating its ability to minimize false positives when classifying benign

breast cancer cases. This strength can be attributed to the ensemble nature of random forests, which combine multiple decision trees to reduce overfitting and improve generalization. This advantage is supported by the study conducted by Baykan et al. (2021), which achieved a precision of 0.89 in breast cancer diagnosis using a random forest classifier. However, the RF classifier faces the challenge of the lowest recall of 0.315 among all models. This indicates a higher risk of misclassifying positive cases as negative, potentially leading to missed diagnoses. This limitation may arise from the inherent bias of decision trees towards majority classes, which can result in underrepresentation of the minority class (benign cases). The study by Chen et al. (2004) discusses this issue. Additionally, the RF classifier exhibits the lowest AUC of 0.806 among all models, suggesting relatively lower overall performance in distinguishing between benign and non-benign cases. This finding aligns with the study by Dittman et al. (2019), which reported lower AUC values for random forest classifiers in breast cancer diagnosis compared to other models.

Lastly, the Support Vector Machine (SVM) classifier also exhibits a notable strength, sharing the highest precision (0.642) among the VGG16 models with the Random Forest classifier. This characteristic suggests its potential to minimize false positives in classifying benign breast cancer, which is crucial in avoiding unnecessary interventions or treatments. The SVM's performance can be attributed to its ability to find the optimal decision boundary that maximizes the margin between classes. This advantage is demonstrated in the study by Huang et al. (2020), which achieved a precision of 0.92 using an SVM classifier for breast cancer diagnosis. However, the SVM classifier has the lowest accuracy of 0.829 and the lowest F1-score of 0.418 among all models. This indicates a potentially higher misclassification rate and a less balanced overall performance. These drawbacks suggest that the SVM classifier may struggle to consistently and reliably perform across different scenarios or data distributions. This limitation could be due to its sensitivity to outliers or the choice of kernel function, as discussed in the study by Duan et al. (2003). In terms of predictive time, the SVM classifier has a runtime of 52.29 seconds, which is faster than the Gradient Boosting classifier but slower than the Random Forest and Logistic Regression classifiers. This difference in speed can be attributed to the optimization process involved in finding the optimal decision boundary, as noted in the study by Steinwart and Christmann (2008).

Based on the preceding analysis, it can be concluded that all classifiers demonstrated highly comparable performance when paired with the VGG16 feature extractor for predicting benign

breast cancer. While the gradient boosting classifier showed a slight advantage in terms of accuracy and AUC metrics, its practical usability is limited by its excessively long prediction time. This extended computation time makes the gradient boosting model less suitable for real-time or time-sensitive clinical applications. In contrast, the logistic regression model emerges as a viable alternative, striking a favorable balance between predictive performance and computational efficiency. Not only did the logistic regression classifier achieve a respectable level of accuracy and AUC, but it also boasted the shortest prediction time among all evaluated models. This combination of reliable predictive capabilities and rapid inference positions the logistic regression model as a compelling choice for deployment in clinical settings, where timely and accurate diagnosis is of utmost importance. Therefore, for clinical applications that require both accurate prediction of benign breast cancer cases and efficient computation, the logistic regression model with the VGG16 feature extractor presents itself as an attractive option. Its balanced trade-off between performance and speed addresses the critical requirements of real-world clinical practice, potentially facilitating more informed decision-making and enhancing patient care.

4.2.1.2 Hybrid AI models based on DenseNet121 feature extractor compared with VGG16 feature extractor in predicting benign breast cancer

Comparing the DenseNet121 feature extractor with the VGG16 feature extractor in predicting benign breast cancer, it is evident that all classifiers exhibit slightly improved performance with DenseNet121. This enhancement can be attributed to the dense connectivity and efficient feature propagation inherent in the DenseNet121 architecture, which have proven highly effective in extracting discriminative features from mammogram images for benign breast cancer prediction. These findings align with the study conducted by Huang et al. (2017), which introduced DenseNet121 and demonstrated its efficacy in various computer vision tasks, including medical image analysis. The dense connectivity pattern within the DenseNet121 model enables the reuse of features and the creation of more compact and informative representations, which is particularly advantageous for complex tasks like breast cancer detection from mammograms. Notably, when paired with the DenseNet121 extractor, the Gradient Boosting (GB) classifier exhibits substantial performance improvements, surpassing its VGG16 counterpart across multiple metrics. With the highest accuracy (0.869) and AUC (0.858) among all models, these gains can be attributed to the ensemble nature of gradient boosting, effectively leveraging the rich and diverse features extracted

by DenseNet121. This finding aligns with Huang et al. (2017), who demonstrated the effectiveness of gradient boosting with deep extractors like DenseNet, achieving an AUC of 0.91. The enhanced precision (0.631) and recall (0.381) indicate an improved ability to distinguish between benign and non-benign cases while minimizing misclassifications. However, the GB classifier with DenseNet121 exhibits a relatively low F1-score (0.508), suggesting a less balanced trade-off between precision and recall. This may be attributed to the inherent complexity of the algorithm and its sensitivity to class imbalance, as noted by Ridgeway (1999). Remarkably, the predictive time of the GB classifier with DenseNet121 is significantly shorter at 187.15 seconds compared to 261.52 seconds with VGG16. This improvement can be attributed to the more efficient feature extraction capabilities of the DenseNet121 architecture, as highlighted by Huang et al. (2017).

The Logistic Regression (LR) classifier, known for its simplicity and interpretability, also demonstrates improved performance with the DenseNet121 extractor. Compared to its VGG16 counterpart, the LR classifier shows enhancements in accuracy (0.856), precision (0.523), F1-score (0.540), and AUC (0.837). These findings align with the study by Xu et al. (2019), which highlights the effectiveness of logistic regression with deep extractors like DenseNet in medical image analysis tasks, achieving an AUC of 0.89. The performance gains can be attributed to the DenseNet121 extractor's ability to capture intricate patterns and relationships within the mammogram data. The linear decision boundary of logistic regression can effectively leverage these representations. Notably, the LR classifier with DenseNet121 achieves the highest F1-score (0.540) among all models, indicating an optimal balance between precision and recall. This aligns with its ability to correctly classify benign and non-benign cases. However, its recall (0.543) is slightly lower than with VGG16 (0.582), potentially due to the linear nature of logistic regression, which may struggle to fully capture the complexity of DenseNet121 representations, as discussed by Dreiseitl and Ohno-Machado (2002). Despite the more complex feature extractor, the LR classifier maintains computational efficiency. With DenseNet121, its predictive time is only slightly longer at 1.14 seconds compared to 0.80 seconds with VGG16. This highlights the LR classifier's computational efficiency, as noted by Le Cessie and Van Houwelingen (1992).

The Random Forest (RF) classifier, renowned for its ability to handle nonlinearities and high-dimensional data, demonstrates considerable performance improvements across all metrics when paired with the DenseNet121 extractor compared to its VGG16 counterpart. Its accuracy (0.853),

precision (0.656), recall (0.342), F1-score (0.513), and AUC (0.836) all surpass those achieved with VGG16, indicating an enhanced capability to accurately classify benign and non-benign cases while minimizing misclassifications. These enhancements stem from the synergy between the ensemble learning approach of random forests and the rich feature representations extracted by DenseNet121. The individual decision trees in the ensemble effectively capture diverse aspects of the complex mammogram data, leading to improved overall performance. These findings align with the study by Baykan et al. (2021), which reported significant improvements in breast cancer diagnosis using random forests with deep extractors like DenseNet, achieving an AUC of 0.92. Notably, the RF classifier with DenseNet121 achieves the highest precision (0.656) among all models, excelling in minimizing false positives, which is crucial for avoiding unnecessary interventions. However, its relatively low recall (0.342) indicates a higher risk of misclassifying positive cases as negative, potentially resulting in missed diagnoses. This limitation can be attributed to the inherent bias of decision trees towards majority classes, further compounded by the complexity of the DenseNet121 representations, as discussed by Chen et al. (2004). Nevertheless, the RF classifier with DenseNet121 exhibits efficient computational performance. Its predictive time of 22.46 seconds is significantly shorter than with VGG16 (36.21 seconds), potentially due to the more efficient feature extraction capabilities of the DenseNet121 architecture, as noted by Liaw and Wiener (2002).

The Support Vector Machine (SVM) classifier, recognized for its robustness and ability to handle high-dimensional data, also demonstrates improved performance when combined with the DenseNet121 extractor compared to its VGG16 counterpart. It achieves higher accuracy (0.847), precision (0.646), F1-score (0.517), and AUC (0.848). These enhancements suggest that the DenseNet121 extractor enables the SVM to better distinguish between benign and non-benign cases while maintaining a relatively high precision, which is crucial for minimizing unnecessary interventions. The performance gains can be attributed to the SVM's effective utilization of the discriminative features extracted by DenseNet121, allowing it to construct an optimal decision boundary in the high-dimensional feature space. These findings align with the study by Huang et al. (2020), which showcases the effectiveness of SVMs with deep extractors like DenseNet in medical image analysis tasks, achieving a precision of 0.92. However, the SVM classifier with DenseNet121 exhibits the lowest accuracy (0.847) and recall (0.332) among the DenseNet121 models. This indicates a higher misclassification rate and the risk of misclassifying positive cases

as negative, potentially leading to missed diagnoses. This limitation may stem from the SVM's sensitivity to the choice of kernel function or the presence of outliers in the DenseNet121 representations, as discussed by Duan et al. (2003). Nonetheless, the SVM classifier remains computationally efficient with the DenseNet121 extractor. Its predictive time of 14.97 seconds is significantly shorter than with VGG16 (52.29 seconds), highlighting its computational efficiency despite the more complex extractor, as noted by Steinwart and Christmann (2008).

The analysis indicates that both VGG16 and DenseNet121, as deep feature extractors, excel in extracting features from mammogram images. However, DenseNet121 exhibits a slight advantage across all classifiers when considering evaluation metrics such as accuracy, precision, recall, F1-score, and AUC. Furthermore, DenseNet121 requires less time for feature extraction compared to VGG16. Among the classifiers, the gradient boosting classifier emerges as the top performer when utilizing both VGG16 and DenseNet121 feature extractors in terms of accuracy and AUC. Nonetheless, it suffers from longer prediction times in both cases. In contrast, logistic regression proves to be a more preferable choice, particularly when combined with DenseNet121. This combination offers a good balance between performance and computational efficiency.

4.2.1.3 Hybrid AI models based on VGG16 feature extractor in predicting malignant breast cancer compared to benign breast cancer

When it comes to predicting malignant breast cancer using original mammogram images, hybrid AI models based on the VGG16 feature extractor demonstrate slight performance improvements. This can be attributed to the inherent characteristics of malignant breast cancer and its internal compositions, which provide more distinct features that enhance the models' predictive capabilities. However, these features can also be complex, potentially impacting the models' performance in an adverse manner. These findings align with a study conducted by Ribli et al. (2018), which explored the characteristics of malignant breast lesions in mammograms and their influence on the performance of deep learning models. The study emphasized that the internal compositions of malignant tumors, such as calcifications, spiculated margins, and architectural distortions, offer valuable discriminative information that can enhance the accuracy of cancer detection. Nevertheless, the complexity of these features can also present challenges for the models, leading to potential variations in performance.

the performance of the proposed classifiers using the VGG16 feature extractor in predicting malignant breast cancer versus benign breast cancer from original mammogram images. Several notable differences emerged from the analysis. The Gradient Boosting (GB) classifier demonstrated slightly lower accuracy (0.856) for malignant prediction compared to benign prediction (0.865). This difference can be attributed to the inherent complexity of detecting malignant cases, which often exhibit more subtle and irregular patterns compared to benign cases. However, the GB classifier displayed higher precision (0.646), recall (0.398), and F1-score (0.480) for malignant prediction, indicating a better balance between minimizing false positives and maximizing true positive detection for malignant cases. These findings are consistent with a study by Harangi et al. (2018), which reported a precision of 0.68 and recall of 0.42 for a gradient boosting classifier using deep features for predicting malignant breast cancer. It is worth noting that the lower AUC (0.822) for malignant prediction suggests a relatively weaker discriminative ability, which could be attributed to the ensemble nature of the GB classifier, making it more susceptible to overfitting or mislabeled instances in the malignant class. Interestingly, the GB classifier exhibited a shorter predictive time (78.38 seconds) for malignant prediction compared to benign prediction (261.52 seconds). This may be a result of the GB classifier's ability to converge faster on the more challenging malignant cases, despite being the longest among the VGG16 models for predicting malignancy.

In contrast, the Logistic Regression (LR) classifier demonstrated higher accuracy (0.852), precision (0.560), F1-score (0.531), and AUC (0.834) for predicting malignant cases compared to benign cases. These findings align with a study by Agarap (2018), which reported an accuracy of 0.86, precision of 0.57, and AUC of 0.86 for a logistic regression classifier using deep features in predicting malignant breast cancer. This indicates that the LR classifier was more effective in capturing the complex patterns associated with malignant cases, resulting in improved overall classification performance and discriminative ability. However, its lower recall (0.460) for malignant prediction, although the highest among the VGG16 models, suggests a higher risk of misclassifying malignant cases as benign, which could be critical in a clinical setting. The LR classifier exhibited a shorter predictive time (0.2 seconds) for malignant prediction compared to benign prediction (0.8 seconds), which is the shortest among the VGG16 models. This highlights the computational efficiency of the LR classifier, potentially attributed to its simpler linear decision boundary.

The Random Forest (RF) classifier displayed superior performance across multiple metrics when predicting malignant cases. It achieved higher accuracy (0.850), precision (0.652 - the highest among VGG16 models), F1-score (0.471), and AUC (0.827). These findings align with a study by Xu et al. (2019), which reported an accuracy of 0.84, precision of 0.67, and AUC of 0.87 for a random forest classifier utilizing deep features in predicting malignant breast cancer. This indicates that the ensemble approach of the RF classifier effectively captured the complex patterns associated with malignant cases while minimizing false positives. However, its lower recall (0.318 - the lowest among all models) for malignant prediction suggests a higher risk of misclassifying malignant cases as benign, which is a significant concern. The RF classifier exhibited a shorter predictive time (10.56 seconds) for malignant prediction compared to benign prediction (36.21 seconds), potentially due to its ability to parallelize computations and converge faster on the more challenging malignant cases.

The Support Vector Machine (SVM) classifier demonstrated higher accuracy (0.846), recall (0.392), F1-score (0.465), and AUC (0.836 - the highest among VGG16 models) when predicting malignant cases compared to benign cases. These findings are consistent with a study by Zheng et al. (2018), which reported an accuracy of 0.85, recall of 0.41, and AUC of 0.88 for an SVM classifier utilizing deep features in predicting malignant breast cancer. This suggests that the SVM classifier's ability to find an optimal separating hyperplane was effective in capturing the complex patterns associated with malignant cases, leading to improved overall classification performance and discriminative ability. However, its lower precision (0.552) for malignant prediction, the lowest among all models, indicates a higher risk of false positives, which could result in unnecessary follow-up procedures or patient anxiety. The SVM classifier exhibited a shorter predictive time (13.42 seconds) for malignant prediction compared to benign prediction (52.29 seconds), potentially due to its ability to converge faster on the more challenging malignant cases, which could be attributed to its structural risk minimization principle.

Upon analysis, it can be concluded that all classifiers, when combined with the VGG16 feature extractor, showed a modest improvement in certain evaluation metrics when predicting malignant breast cancer compared to benign breast cancer. The gradient boosting classifier stood out as the top-performing model among the evaluated classifiers, achieving the highest accuracy score. However, this increased accuracy came at the expense of a lower area under the receiver operating

characteristic curve (AUC) value compared to its performance in predicting benign breast cancer cases. Furthermore, the gradient boosting classifier exhibited a longer prediction time, limiting its practical usability. On the other hand, the logistic regression classifier emerged as a viable alternative, surpassing the gradient boosting classifier in terms of the AUC metric while also demonstrating a shorter prediction time. This combination of competitive discriminative ability, as indicated by the AUC, and computational efficiency makes the logistic regression classifier a more favorable choice for practical applications involving the prediction of malignant breast cancer from mammogram images when utilizing the VGG16 feature extractor.

4.2.1.4 Hybrid AI models based on DenseNet121 feature extractor compared with VGG16 feature extractor in predicting malignant breast cancer

Upon further analysis, it is evident that the utilization of the DenseNet121 feature extractor resulted in improved performance across all classifiers in predicting malignant breast cancer cases. This improvement aligns with previous findings where the DenseNet121 feature extractor outperformed the VGG16 feature extractor in predicting benign breast cancer cases. This trend, illustrating the comparative performance of the classifiers when coupled with the DenseNet121 and VGG16 feature extractors for predicting malignant breast cancer instances. The consistent pattern observed in both benign and malignant cancer prediction scenarios highlights the effectiveness of the DenseNet121 architecture in extracting informative features from mammogram images. The ability of this feature extractor to capture more meaningful representations of the input data likely contributed to the enhanced classification performance exhibited by all evaluated models when identifying cases of malignant breast cancer. This discovery emphasizes the crucial role of feature extraction techniques in developing accurate breast cancer prediction systems. By leveraging advanced architectures such as DenseNet121, which excel at extracting relevant features from complex medical imaging data, it is possible to enhance the diagnostic capabilities of various classification algorithms, ultimately leading to more precise and reliable predictions for both benign and malignant breast cancer cases.

When predicting cases of malignant breast cancer, the Logistic Regression (LR) classifier demonstrated higher accuracy, recall, F1-score, and AUC when coupled with the DenseNet121 feature extractor as compared to the VGG16 feature extractor. This improvement can be attributed to the enhanced representational capabilities of the DenseNet121 architecture, which enabled it to

effectively capture the intricate patterns associated with malignant tumors in mammogram images. The dense connectivity and feature reuse mechanisms within DenseNet121 likely facilitated the learning of more comprehensive and distinctive features, thereby contributing to the improved performance in discriminating malignant cases. However, the lower precision of the LR classifier with DenseNet121 suggests a higher likelihood of misclassifying benign cases as malignant, potentially leading to unnecessary biopsies or additional testing. Nevertheless, the significantly shorter predictive time achieved with DenseNet121 (0.1 seconds) compared to VGG16 (0.2 seconds) underscores its computational efficiency, rendering it more suitable for time-sensitive clinical applications.

Although the Gradient Boosting (GB) classifier exhibited higher accuracy and recall when using the DenseNet121 feature extractor for malignant prediction compared to VGG16, its lower precision, F1-score, and AUC indicate a less optimal balance between correctly identifying true positives and minimizing false positives. The ensemble nature of the GB classifier, combined with the increased complexity of the DenseNet121 feature extractor, might have contributed to overfitting, resulting in a higher rate of false positive predictions. Intriguingly, the GB classifier's AUC for malignant cases was lower with DenseNet121 than for benign cases, indicating that the model faced greater difficulty in distinguishing malignant patterns compared to benign ones, possibly due to the inherent complexity and variability of malignant tumors. Additionally, the significantly longer predictive time with DenseNet121 (122.78 seconds) compared to VGG16 (78.38 seconds) could be seen as a trade-off for the heightened model complexity, potentially limiting its practical usability in time-sensitive scenarios.

The Random Forest (RF) classifier demonstrated improved accuracy, precision, F1-score, and AUC when utilizing the DenseNet121 feature extractor compared to VGG16 for predicting malignant cases. This indicates enhanced overall classification performance and better control over false positive rates. The ensemble nature of the RF classifier, combined with the powerful DenseNet121 feature extractor, likely facilitated effective learning of the intricate patterns associated with malignant tumors while maintaining a favorable balance between precision and recall. However, the lower recall observed with DenseNet121 suggests an increased risk of misclassifying malignant cases as benign, which could have significant implications in a clinical setting. Despite this, the longer predictive time with DenseNet121 (15.58 seconds) compared to

VGG16 (10.56 seconds) may be considered a trade-off for the heightened complexity of the model, yet it remains relatively efficient compared to other classifiers.

When utilizing the DenseNet121 feature extractor, the Support Vector Machine (SVM) classifier demonstrated higher precision, recall, F1-score, and AUC compared to VGG16 for predicting malignant cases. This indicates an overall improvement in classification performance, discriminative ability, and a better balance between minimizing false positives and maximizing true positive detection. The SVM's capability to determine an optimal decision boundary in high-dimensional spaces, combined with the powerful DenseNet121 feature extractor, likely facilitated effective differentiation between malignant and benign cases. However, the lower accuracy observed with DenseNet121 compared to VGG16 suggests a higher overall misclassification rate, which could be a concern in clinical applications. It is worth noting that the shorter predictive time achieved with DenseNet121 (11.71 seconds) compared to VGG16 (13.42 seconds) can be attributed to the SVM classifier's ability to converge more quickly with the more discriminative DenseNet121 features. This potentially offsets the increased complexity of the model.

After conducting a thorough analysis of the aforementioned results, several key points emerge regarding the proposed hybrid models for breast cancer prediction from mammogram images. These points shed light on the performance and efficacy of the models. Firstly, the hybrid models displayed commendable performance in predicting both benign and malignant breast cancer cases, with a slight advantage in predicting malignant cases. This highlights the models' ability to effectively capture important patterns and features indicative of breast cancer. Deep learning techniques proved to be highly effective in extracting informative features from mammogram images, a critical step in achieving accurate cancer prediction. Specifically, the DenseNet121 feature extractor exhibited marginal superiority over the VGG16 feature extractor in terms of the discriminative power of the extracted features. This demonstrates the importance of selecting an appropriate deep learning architecture for feature extraction. Among the evaluated classifiers, the gradient boosting model initially demonstrated strong classification capabilities. However, its performance gradually declined as the complexity of the task increased. Ultimately, the logistic regression classifier outperformed the gradient boosting model when paired with the DenseNet121 feature extractor for predicting malignant breast cancer cases. The logistic regression classifier not only exhibited competitive predictive performance but also boasted a shorter prediction time

compared to other model configurations. This reduced computational burden enhances the practicality of deploying the logistic regression classifier in clinical applications where timely and efficient diagnoses are crucial. These collective findings underscore the promising potential of hybrid models that combine deep learning-based feature extraction techniques with traditional machine learning classifiers for accurate breast cancer prediction from mammogram images. The superior feature extraction capabilities of the DenseNet121 architecture, combined with the logistic regression classifier's balanced performance and computational efficiency, make this combination a compelling approach for real-world clinical deployments. It offers both diagnostic accuracy and practical feasibility, providing a valuable tool for breast cancer diagnosis and treatment.

4.2.2 Training Stage 2: From Enhanced Mammogram Images

The results from the previous experiments demonstrate a notable enhancement in the performance of all the hybrid models proposed when image enhancement techniques are applied, as compared to the performance observed with the original, unenhanced images. Various techniques, including morphological erosion preprocessing, contrast-limited adaptive histogram equalization (CLAHE), Laplacian of Gaussian (LoG) edge enhancement, and non-sharp masking, were employed to enhance the images and yielded significant improvements in the models' capabilities. Specifically, the hybrid models achieved an impressive accuracy range of 0.952 to 0.991 for benign breast cancer prediction and an even higher accuracy range of 0.959 to 0.995 for malignant breast cancer prediction. This remarkable improvement can be attributed to the crucial role played by image enhancement techniques in enhancing the clarity and visual quality of mammogram images.

By utilizing techniques like CLAHE to enhance contrast or unsharp masking to sharpen edges and details, the resulting enhanced images become more suitable for feature extractors to capture discriminative and informative features. When these enhanced images are fed into the deep learning component of the hybrid models, feature extractors such as DenseNet121 and VGG16 can more effectively learn and extract relevant patterns and representations necessary for accurate breast cancer prediction. The improved image quality resulting from the enhancement techniques allows the feature extractors to capture finer details, subtle variations, and crucial characteristics that may have been overlooked or obscured in the original unenhanced images. This finding is further supported by a study conducted by Sharma et al. (2020), which demonstrated the positive

impact of image enhancement techniques on the performance of deep learning models for breast cancer detection from mammograms. The study reported an increase in accuracy from 92.1% to 94.3% when CLAHE and unsharp masking were employed as preprocessing steps before feeding the images to a deep learning model. This improvement in accuracy can be attributed to the enhanced image quality, which facilitated the deep learning model's ability to learn more discriminative features for accurate breast cancer detection.

4.2.2.1 Hybrid AI models based on VGG16 feature extractor after enhancement compared with before enhancement in predicting benign breast cancer

The Logistic Regression (LR) classifier with the VGG16 feature extractor exhibited remarkable improvements across all evaluation metrics after applying image enhancement techniques. The accuracy increased from 0.848 to an impressive 0.985, precision surged from 0.511 to 0.986, recall soared from 0.582 to an outstanding 0.979, F1-score rose from 0.524 to 0.987, and the Area under the Receiver Operating Characteristic Curve (AUC) climbed from 0.814 to a remarkable 0.987. These substantial gains can be attributed to the enhanced contrast, edge details, and noise reduction provided by the image enhancement techniques. Consequently, the VGG16 feature extractor captured the salient characteristics of benign breast tissue patterns more effectively, thereby amplifying the LR classifier's ability to discriminate between benign and malignant cases. Notably, the predictive time for the LR classifier decreased significantly from 0.80 seconds to a mere 0.13 seconds, highlighting increased computational efficiency and suitability for time-sensitive clinical applications.

Similarly, the Support Vector Machine (SVM) classifier with the VGG16 feature extractor demonstrated significant performance improvements across all metrics after employing image enhancement techniques. The accuracy rose from 0.829 to 0.975, precision increased from 0.642 to 0.981, recall skyrocketed from 0.319 to 0.959, F1-score doubled from 0.418 to 0.841, and the AUC soared from 0.834 to 0.986. The improved performance can be attributed to the SVM's ability to find an optimal decision boundary in the enhanced feature space provided by the VGG16 feature extractor and image enhancement techniques. This facilitated more effective separation of benign cases. Additionally, the predictive time for the SVM classifier decreased significantly from 52.29 seconds to 4.45 seconds, contributing to improved efficiency in clinical settings.

The application of image enhancement techniques improved the performance of the Gradient Boosting (GB) classifier with the VGG16 feature extractor in predicting benign breast cancer. The accuracy increased from 0.865 to 0.973, recall rose from 0.338 to 0.867, F1-score more than doubled from 0.432 to 0.924, and the AUC soared from 0.841 to an impressive 0.987, the highest among all VGG16 models. Although the precision of 0.978 was lower than that of the LR and SVM classifiers, it was still substantially higher than the pre-enhancement value of 0.854. The ensemble nature of the GB classifier, combined with the enhanced image features, likely contributed to improved learning of benign patterns. Notably, the predictive time for the GB classifier decreased from 261.52 seconds to 58.25 seconds, although it remained the longest among the VGG16 models.

The Random Forest (RF) classifier with the VGG16 feature extractor also demonstrated improved performance after applying image enhancement techniques. The accuracy increased from 0.846 to 0.952, precision rose from 0.610 to 0.985, recall increased from 0.315 to 0.803, F1-score improved from 0.460 to 0.814, and the AUC climbed from 0.806 to 0.976. However, it had the lowest accuracy, recall, F1-score, and AUC among all models after enhancement, indicating room for further improvement. The ensemble nature of the RF classifier, combined with the enhanced image features, likely contributed to better learning of benign patterns, although its performance may have been limited by the inherent complexity of the VGG16 feature extractor. Nonetheless, the shorter predictive time of 4.83 seconds compared to 36.21 seconds for pre-enhanced images is noteworthy.

While image enhancement techniques consistently improved the performance of all classifiers, there were trade-offs between different performance metrics and computational efficiency across the models. The Logistic Regression classifier exhibited the highest overall performance, with near-perfect scores across all metrics and the shortest predictive time, making it an outstanding choice for benign prediction in time-sensitive clinical settings. The Support Vector Machine classifier closely followed, with excellent performance and a notable reduction in predictive time. The Gradient Boosting classifier demonstrated remarkable discriminative ability (AUC), but its lower precision compared to other models should be taken into account. The Random Forest classifier had the lowest overall performance, suggesting it may not be the most suitable choice for this task. Ultimately, the selection of the most appropriate model should consider specific

requirements and priorities, such as minimizing false positives, maximizing true positive detection, achieving a balanced trade-off between precision and recall, or prioritizing computational efficiency in a clinical setting.

4.2.2.2 Hybrid AI models based on DenseNet121 feature extractor compared with VGG16 feature extractor after enhancement in predicting benign breast cancer

After applying image enhancement techniques, the hybrid AI models using the DenseNet121 feature extractor outperformed the models using the VGG16 feature extractor in predicting benign breast cancer. However, the trade-offs between evaluation metrics and computational efficiency varied among the classifiers, emphasizing the need to consider specific requirements when selecting the most suitable model. The Logistic Regression (LR) classifier with the DenseNet121 feature extractor achieved the highest overall performance among all models. It exhibited improved accuracy (0.991 vs. 0.985 with VGG16), precision (0.996 vs. 0.986 with VGG16), F1-score (0.989 vs. 0.987 with VGG16), and AUC (0.999 vs. 0.987 with VGG16). Although its recall (0.978) was slightly lower than the VGG16 model (0.979), it remained the highest among the DenseNet121 models. With near-perfect scores across all metrics and the shortest predictive time (0.15 seconds vs. 0.13 seconds with VGG16) among DenseNet121 models, the LR classifier with DenseNet121 is an excellent choice for benign prediction in time-sensitive clinical settings. The enhanced discriminative power of DenseNet121 features, combined with the simplicity of the logistic regression model, likely contributed to its exceptional performance.

The Support Vector Machine (SVM) classifier with the DenseNet121 feature extractor demonstrated the second-highest performance. It achieved the highest AUC (0.999, tied with the LR classifier) and the second-highest accuracy (0.986), precision (0.991), and F1-score (0.987) among all models. However, its recall (0.946) was lower than the LR classifier with DenseNet121 and the VGG16 models. The shorter predictive time (8.08 seconds) compared to the pre-enhanced model and the VGG16 model (4.45 seconds) is noteworthy. The SVM's ability to find an optimal decision boundary in the enhanced feature space provided by DenseNet121, along with the image enhancement techniques, likely contributed to its strong performance.

The Random Forest (RF) classifier with the DenseNet121 feature extractor exhibited improved performance compared to the pre-enhanced and VGG16 models, with higher accuracy (0.969 vs. 0.952 with VGG16), recall (0.881 vs. 0.803 with VGG16), and AUC (0.994 vs. 0.976 with

VGG16). However, its precision (0.902) was the lowest among all models, indicating a higher risk of false positive predictions for benign cases. Additionally, its F1-score (0.836) was the lowest among DenseNet121 models, suggesting lower overall performance in discriminating benign cases compared to other classifiers. The shorter predictive time (15.43 seconds) compared to the pre-enhanced model and the VGG16 model (4.83 seconds) is noteworthy. The ensemble nature of the RF classifier, combined with the enhanced image features provided by DenseNet121, likely facilitated improved learning of benign patterns, although its performance may have been limited by the inherent complexity of the DenseNet121 feature extractor.

The DenseNet121 feature extractor, combined with image enhancement techniques, yielded performance improvements across most classifiers compared to the VGG16 feature extractor. The Logistic Regression classifier demonstrated the highest overall performance, closely followed by the Support Vector Machine classifier. The Random Forest classifier exhibited improved performance but had the lowest precision and F1-score among all models. The Gradient Boosting classifier showed enhanced performance compared to the pre-enhanced and VGG16 models, yet it had the lowest accuracy, recall, and AUC among the DenseNet121 models. While the GB classifier benefited from the ensemble nature and enhanced image features provided by DenseNet121, there is room for improvement in terms of accuracy, recall, and AUC. It's worth noting that the GB classifier had the longest predictive time among all models, albeit shorter than the pre-enhanced model, which may limit its practicality in time-sensitive scenarios. When selecting the most suitable model, it is crucial to consider specific requirements and priorities, such as minimizing false positives, maximizing true positive detection, achieving a balanced trade-off between precision and recall, or prioritizing computational efficiency in a clinical setting.

4.2.2.3 Hybrid AI models based on VGG16 feature extractor after enhancement in predicting malignant breast cancer compared to benign breast cancer

The Logistic Regression (LR) classifier with the VGG16 feature extractor demonstrated the highest overall performance among VGG16 models in predicting malignancy. It achieved the highest accuracy (0.983, compared to 0.852 for pre-enhanced images), F1-score (0.955, compared to 0.531 for pre-enhanced), and Area under the Receiver Operating Characteristic Curve (AUC) (0.996, compared to 0.834 for pre-enhanced). These improvements can be attributed to the enhanced discriminative power of VGG16 features combined with the simplicity of the logistic

regression model in distinguishing malignant patterns. Despite its lower precision (0.960), indicating a higher risk of false positives, the LR classifier had the shortest predictive time (0.08 seconds) among all models, making it suitable for time-sensitive clinical applications.

The Support Vector Machine (SVM) classifier with the VGG16 feature extractor achieved the second-highest accuracy (0.980, compared to 0.846 for pre-enhanced images) among VGG16 models in predicting malignancy. It exhibited the highest precision (0.989) and recall (0.961) among VGG16 models, indicating excellent performance in correctly identifying malignant cases. The high AUC (0.994, compared to 0.836 for pre-enhanced) further highlighted the SVM's strong discriminative ability. However, its F1-score (0.848) was the lowest among all models, suggesting an imbalance between precision and recall that should be considered based on specific clinical requirements. The SVM classifier demonstrated a significantly shorter predictive time (3.74 seconds) compared to the pre-enhanced model, making it more computationally efficient than the Gradient Boosting and Random Forest classifiers. The SVM's ability to find an optimal decision boundary in the enhanced feature space provided by VGG16, along with the image enhancement techniques, likely contributed to its strong performance.

The Gradient Boosting (GB) classifier with the VGG16 feature extractor exhibited improved performance compared to pre-enhanced images. It achieved higher accuracy (0.976, compared to 0.856 for pre-enhanced), precision (0.977, compared to 0.646 for pre-enhanced), recall (0.874, compared to 0.398 for pre-enhanced), and F1-score (0.933, compared to 0.480 for pre-enhanced). Its AUC (0.991, compared to 0.822 for pre-enhanced) also increased, indicating strong discriminative ability. However, its AUC was the lowest among all models, indicating room for further improvement. The GB classifier had the longest predictive time (43.94 seconds) among VGG16 models, although it was shorter than the pre-enhanced model, which may limit its practicality in time-sensitive scenarios. The ensemble nature of the GB classifier, combined with the enhanced image features provided by VGG16, likely contributed to improved learning of malignant patterns, but the trade-off between different performance metrics should be carefully considered.

The Random Forest (RF) classifier with the VGG16 feature extractor exhibited the lowest accuracy (0.959, compared to 0.850 for pre-enhanced images) among all models for malignant prediction. However, it achieved the highest precision (0.988) among all models, indicating excellent

performance in correctly identifying malignant cases. Its recall (0.825) was also the highest among all models, suggesting a lower risk of missing true malignant cases, which is crucial in clinical settings. Additionally, the RF classifier had high F1-score (0.893) and AUC (0.992, compared to 0.827 for pre-enhanced), indicating strong overall performance in discriminating malignant cases. Its predictive time (5.33 seconds) was shorter than the pre-enhanced model and the Gradient Boosting classifier, making it more computationally efficient than the latter. The ensemble nature of the RF classifier, combined with the enhanced image features provided by VGG16, likely contributed to improved learning of malignant patterns, although its lower accuracy compared to other models should be considered within the context of specific clinical priorities.

In summary, while all classifiers demonstrated improved performance in predicting malignancy using the VGG16 feature extractor and image enhancement techniques, the Logistic Regression classifier exhibited the highest overall performance among VGG16 models. It achieved the highest accuracy, F1-score, and AUC, and had the shortest predictive time. The Support Vector Machine classifier closely followed, with the highest precision and recall among VGG16 models, albeit with a lower F1-score. The Gradient Boosting classifier showed improved performance but had the lowest AUC among all models and the longest predictive time among VGG16 models. The Random Forest classifier demonstrated the highest precision and recall among all models but the lowest accuracy. When selecting the most appropriate model, careful consideration should be given to specific requirements and priorities, such as minimizing false positives or false negatives, achieving a balanced trade-off between precision and recall, or prioritizing computational efficiency in a clinical setting.

4.2.2.4 Hybrid AI models based on DenseNet121 feature extractor compared with VGG16 feature extractor after enhancement in predicting malignant breast cancer

The Logistic Regression (LR) classifier with the DenseNet121 feature extractor achieved the highest overall performance for malignant prediction. It demonstrated the highest accuracy (0.995, compared to 0.983 with VGG16), precision (0.995, compared to 0.960 with VGG16), recall (0.995, compared to 0.950 with VGG16), F1-score (0.995, compared to 0.955 with VGG16), and AUC (0.999, compared to 0.996 with VGG16) among all models. These exceptional scores, combined with the shortest predictive time (0.22 seconds, compared to 0.08 seconds with VGG16) among DenseNet121 models, make the LR classifier with DenseNet121 an outstanding choice for

malignant prediction in time-sensitive clinical settings. The enhanced discriminative power of DenseNet121 features, combined with the simplicity of the logistic regression model, likely contributed to its remarkable performance in distinguishing malignant patterns from benign ones.

The Support Vector Machine (SVM) classifier with the DenseNet121 feature extractor demonstrated the second-highest performance among all models. It achieved the second-highest accuracy (0.992), precision (0.992), F1-score (0.982), and AUC (0.998) among all models. Its recall (0.966) was slightly lower than the LR classifier with DenseNet121 but higher than the VGG16 models. The substantial improvements across all metrics compared to pre-enhanced images and the VGG16 feature extractor highlight the SVM's ability to leverage the enhanced feature space provided by DenseNet121, combined with the image enhancement techniques. However, its longer predictive time (19.03 seconds) compared to the LR classifier and the VGG16 models should be considered in time-sensitive applications.

The Gradient Boosting (GB) classifier with the DenseNet121 feature extractor demonstrated enhanced performance compared to pre-enhanced images and the VGG16 feature extractor. It achieved higher accuracy (0.982, compared to 0.976 with VGG16), precision (0.979, compared to 0.977 with VGG16), recall (0.892, compared to 0.874 with VGG16), F1-score (0.947, compared to 0.933 with VGG16), and AUC (0.996, compared to 0.991 with VGG16). However, its precision, recall, F1-score, and AUC were the lowest among DenseNet121 models, indicating potential for further improvement. It is worth noting that the GB classifier had the longest predictive time (219.23 seconds) among all models, although it was shorter than the pre-enhanced model. This may limit its practical usefulness in time-sensitive scenarios. The ensemble nature of the GB classifier, combined with the enhanced image features provided by DenseNet121, likely contributed to improved learning of malignant patterns. However, the trade-off between different performance metrics should be carefully considered.

The Random Forest (RF) classifier with the DenseNet121 feature extractor demonstrated lower accuracy (0.973) and AUC (0.996) compared to other DenseNet121 models, although it improved compared to pre-enhanced images and the VGG16 feature extractor. However, it achieved the second-highest precision (0.988) among all models, indicating excellent performance in correctly identifying malignant cases. Its recall (0.895) and F1-score (0.950) were lower than the LR and SVM classifiers with DenseNet121 but higher than the VGG16 models. The predictive time (25.49

seconds) was longer than the LR and SVM classifiers but shorter than the Gradient Boosting classifier. The ensemble nature of the RF classifier, combined with the enhanced image features provided by DenseNet121, likely contributed to improved learning of malignant patterns. However, its lower accuracy and AUC compared to other DenseNet121 models should be considered in the context of specific clinical priorities.

In summary, the DenseNet121 feature extractor, along with image enhancement techniques, resulted in substantial performance improvements across all classifiers for predicting malignant cases compared to pre-enhanced images and the VGG16 feature extractor. The Logistic Regression classifier demonstrated the highest overall performance, closely followed by the Support Vector Machine classifier. The Gradient Boosting classifier exhibited improved performance but had the lowest precision, recall, F1-score, and AUC among DenseNet121 models, as well as the longest predictive time. The Random Forest classifier showed the second-highest precision among all models but had the lowest accuracy and AUC among DenseNet121 models. When selecting the most appropriate model, specific requirements and priorities, such as minimizing false positives or false negatives, achieving a balanced trade-off between precision and recall, or prioritizing computational efficiency in a clinical setting, should be taken into consideration.

4.2.3 Application Stage: From New Mammogram Images

In the application stage, where new mammogram images are used, the performance of all proposed hybrid models showed a slight decrease compared to their performance on the training data. The accuracy range for predicting benign breast cancer was between 0.833 and 0.967, while the accuracy range for predicting malignant breast cancer was between 0.901 and 0.955. This decrease in performance can be attributed to a phenomenon known as domain shift or dataset shift. Domain shift occurs when the distribution of the data used for training differs from the distribution of the data encountered during testing or deployment. In the context of mammogram image analysis, the training data may not fully capture the characteristics or patterns present in real-world clinical settings. Factors such as differences in imaging equipment, patient demographics, or tumor characteristics across healthcare facilities or geographic regions can contribute to domain shift. When the hybrid models are exposed to new, unseen images that deviate from the distribution of the training data, their performance may decrease slightly as they encounter patterns or variations that were not adequately represented during training.

A study conducted by Zhu et al. (2020) supports this finding, indicating that deep learning models for breast cancer detection from mammograms experience a decrease in performance when tested on data from different sources or healthcare facilities. This suggests that domain shift is a common challenge when deploying these models in real-world clinical settings. To mitigate the effects of domain shift and improve the generalization ability of the hybrid models, several strategies can be employed. These include data augmentation techniques, transfer learning, domain adaptation techniques, or incorporating more diverse and representative data during the training process. By addressing the issue of domain shift, the performance of the hybrid models can be further optimized, ensuring more consistent and reliable predictions when applied to new, unseen mammogram images in clinical practice.

4.2.3.1 Hybrid AI models based on VGG16 feature extractor post-application compared with pre-application in predicting benign breast cancer

The Logistic Regression (LR) classifier with the VGG16 feature extractor showed a significant decrease in accuracy from 0.985 to 0.942 and precision from 0.986 to 0.872 when applied to the new mammogram dataset. This drop in performance is attributed to domain shift, where the distribution of the new dataset differs from the training data. However, the LR classifier maintained a high prediction percentage of 94.0% and recall of 0.944, indicating its ability to correctly identify a large proportion of positive (malignant) cases. The F1-score decreased from 0.987 to 0.907 but remained the highest among the VGG16 models. The AUC slightly dipped from 0.987 to 0.986, suggesting that the LR classifier's ability to discriminate between benign and malignant cases remained reasonably robust. Notably, it had the fastest predictive time of 0.05 seconds among all models, making it suitable for time-sensitive clinical applications.

The Gradient Boosting (GB) classifier with VGG16 experienced a decrease in accuracy from 0.973 to 0.925, and its recall dropped sharply from 0.867 to 0.777, indicating a higher rate of false negatives on the new dataset. However, it maintained the highest precision of 0.965 among all models, indicating a lower rate of false positives. The F1-score declined from 0.924 to 0.862, and the AUC reduced from 0.987 to 0.947, indicating a decrease in overall performance. This classifier had the longest predictive time of 23.30 seconds among the VGG16 models, which may limit its practical utility in clinical settings.

The Support Vector Machine (SVM) classifier with VGG16 exhibited a drop in accuracy from 0.975 to 0.916, and its precision decreased to 0.722, the lowest among all models. This low precision could result in a higher rate of false positives on the new dataset, potentially leading to unnecessary follow-up procedures or patient anxiety. Recall fell from 0.959 to 0.893, but the F1-score remained relatively steady, decreasing slightly from 0.841 to 0.838. Interestingly, the AUC marginally increased from 0.986 to 0.988, the highest among the VGG16 models, indicating overall good performance despite the decline in other metrics. The predictive time decreased from 4.45 seconds to 1.78 seconds, which could be advantageous in clinical settings.

The Random Forest (RF) classifier with VGG16 showed the most significant decline in performance when applied to the new dataset. Its accuracy dropped from 0.952 to 0.833, the lowest among all models, and the prediction percentage was also the lowest at 90.4%. Precision decreased from 0.985 to 0.958, and recall sharply declined from 0.803 to 0.638, the lowest across all models. The F1-score decreased from 0.814 to 0.766, also the lowest, and the AUC declined steeply from 0.976 to 0.930, the lowest among the models. This substantial performance drop for the RF classifier indicates a lack of robustness to domain shift, possibly due to its tendency to overfit to the training data.

These findings align with a study conducted by Zhu et al. (2020), which observed a decrease in performance when deep learning models for breast cancer detection were tested on data from different sources or healthcare facilities, indicating the presence of domain shift. Additionally, Kyono et al. (2019) found in their study that certain models, such as ensemble methods like gradient boosting, showed better generalization performance and were more resilient to domain shift compared to other models like random forests. Furthermore, Sharma et al. (2021) investigated domain adaptation techniques to mitigate the effects of domain shift in mammogram classification. Their study demonstrated that applying domain adaptation methods, such as adversarial domain adaptation and instance weighting, could improve the performance of deep learning models on new, unseen datasets, reducing the impact of domain shift.

4.2.3.2 Hybrid AI models based on DenseNet121 feature extractor compared with VGG16 post-application in predicting benign breast cancer

Comparing the performance of hybrid AI models based on the DenseNet121 feature extractor with post-application to the VGG16 feature extractor in predicting benign breast cancer, we observed

that the DenseNet121 models showed better results. The Logistic Regression (LR) classifier with the DenseNet121 feature extractor demonstrated strong performance, achieving the highest accuracy of 0.967 among all models. It also had the maximum prediction percentage of 95.2% and the highest recall of 0.975. Although its precision decreased from 0.996 to 0.886, the F1-score remained the highest at 0.939. Remarkably, the AUC remained unchanged at 0.999, indicating that its ability to distinguish between benign and malignant cases was unaffected by domain shift. Additionally, this classifier exhibited the fastest predictive time of 0.06 seconds among all DenseNet121 models, making it well-suited for clinical applications.

The Random Forest (RF) classifier with DenseNet121 observed a decrease in accuracy from 0.969 to 0.950, but it achieved the highest prediction percentage of 95.2%. While precision improved from 0.902 to 0.933, recall decreased from 0.881 to 0.824. The F1-score slightly increased from 0.836 to 0.903, while the AUC declined from 0.994 to 0.984, indicating a slight decrease in overall performance. The Gradient Boosting (GB) classifier with DenseNet121 experienced a decline in accuracy from 0.957 to 0.933, the lowest among DenseNet121 models. The prediction percentage of 92.9% was also the lowest. Although precision remained the highest among DenseNet121 models, recall dropped from 0.872 to 0.794, and the F1-score decreased from 0.939 to 0.871. The AUC declined from 0.992 to 0.973, the lowest among DenseNet121 models, suggesting a decrease in overall performance. However, this classifier had the longest predictive time of 56.42 seconds among all models, which may limit its practical utility in clinical settings.

The Support Vector Machine (SVM) classifier with DenseNet121 exhibited a decline in accuracy from 0.986 to 0.933, the lowest among DenseNet121 models. Its precision dropped significantly from 0.991 to 0.814, the lowest across all models, indicating a higher rate of false positives on the new dataset. Recall also sharply decreased from 0.946 to 0.765, the lowest among DenseNet121 models, indicating a higher rate of false negatives. Consequently, the F1-score decreased from 0.987 to 0.866, also the lowest among DenseNet121 models. The AUC declined from 0.999 to 0.980, but the predictive time improved from 8.08 seconds to 3.23 seconds.

These findings align with previous studies mentioned, highlighting the impact of domain shift on the performance of machine learning models, including deep learning-based approaches, when applied to new, and unseen datasets. The studies conducted by Zhu et al. (2020) and Kyono et al. (2019) confirm that the performance of breast cancer detection models can decrease when tested

on data from different sources or healthcare facilities due to the presence of domain shift. Although the extent of performance decline varied across classifiers and metrics, the Logistic Regression (LR) classifier with both the VGG16 and DenseNet121 feature extractors demonstrated the most robust performance, maintaining high accuracy, recall, and AUC values. This suggests that the combination of deep learning-based feature extraction and logistic regression classification may be more resilient to domain shift compared to other classifiers such as Random Forest, Gradient Boosting, or Support Vector Machines.

4.2.3.3 Hybrid AI models based on VGG16 feature extractor post-application in predicting malignant breast cancer compared to benign breast cancer

Comparing the performance of hybrid AI models based on the VGG16 feature extractor in predicting malignant breast cancer versus benign breast cancer, we observed that the classifiers exhibited slightly better results for malignant breast cancer prediction. The Logistic Regression (LR) classifier with the VGG16 feature extractor experienced a decrease in accuracy from 0.983 to 0.937, but it remained the highest among VGG16 models. Its prediction percentage of 94.4% was respectable, and although precision fell from 0.960 to 0.902, it was still the highest among VGG16 extractors. Recall also declined from 0.950 to 0.925, yet it remained the highest, indicating a low rate of false negatives. The F1-score decreased from 0.955 to 0.914 but remained the highest among VGG16 models. The AUC dipped from 0.996 to 0.994 but remained the highest for this extractor, suggesting good overall performance. Notably, the predictive time decreased from 0.08 seconds to 0.06 seconds, the shortest among all models, making it well-suited for time-sensitive clinical applications.

The Random Forest (RF) classifier with VGG16 observed a drop in accuracy from 0.959 to 0.923 after post-enhancement. Although the prediction percentage of 93.1% was respectable, precision plunged from 0.988 to 0.816, indicating a higher rate of false positives. Recall also decreased from 0.825 to 0.800, and the F1-score declined from 0.893 to 0.811. Importantly, the AUC sharply decreased from 0.992 to 0.948, the lowest among all models, suggesting a significant decrease in overall performance. For the Gradient Boosting (GB) classifier with VGG16, accuracy declined from 0.976 to 0.921 after post-enhancement. The prediction percentage of 91.7% was the lowest among all models. While precision dipped from 0.977 to 0.968, it remained the highest overall, indicating a low rate of false positives. However, recall experienced a steep drop from 0.874 to

0.750, the lowest for VGG16 extractors, suggesting a higher rate of false negatives. The F1-score decreased from 0.933 to 0.845, and the AUC reduced from 0.991 to 0.974. The accuracy of the Support Vector Machine (SVM) classifier with VGG16 significantly dropped from 0.980 to 0.901, which was the lowest among all models. Its prediction percentage tied for the lowest at 91.7%. Additionally, there was a concerning decline in precision from 0.989 to 0.706, the lowest overall, indicating a high rate of false positives. Recall also decreased from 0.961 to 0.852. The F1-score slightly declined from 0.848 to 0.810, the lowest across all classifiers. The AUC dipped from 0.994 to 0.991.

These findings are consistent with the previously mentioned studies, emphasizing the impact of domain shift on the performance of machine learning models, including deep learning-based approaches, when applied to new, unseen datasets. The studies conducted by Zhu et al. (2020) and Kyono et al. (2019) confirm that the performance of breast cancer detection models can decrease when tested on data from different sources or healthcare facilities due to the presence of domain shift. Similar to the observations made for benign prediction, the Logistic Regression (LR) classifier with the VGG16 feature extractor demonstrated the most robust performance for malignant prediction, maintaining high accuracy, recall, and AUC values. This further underscores the potential resilience of the combination of deep learning-based feature extraction and logistic regression classification to domain shift.

4.2.3.4 Hybrid AI models based on DenseNet121 feature extractor compared with VGG16 feature extractor post-application in predicting malignant breast cancer

Comparing the performance of hybrid AI models based on the DenseNet121 feature extractor with the VGG16 feature extractor in predicting malignant breast cancer, we observed slightly lower results for the classifiers when utilizing DenseNet121. The Logistic Regression (LR) classifier with the DenseNet121 feature extractor demonstrated strong performance despite image enhancement. Although its accuracy decreased from 0.995 to 0.955, it remained higher than all VGG16 models. The prediction percentage of 95.8% was the highest overall, and precision decreased from 0.995 to 0.921 but remained the highest among all classifiers. Recall declined from 0.995 to 0.946 while still being the highest, indicating a low rate of false negatives. The F1-score reduced from 0.995 to 0.933 but remained the highest, and the AUC slightly dipped from 0.999 to 0.998 while maintaining the highest value overall, suggesting excellent discriminative ability.

The Gradient Boosting (GB) classifier with DenseNet121 experienced a decrease in accuracy from 0.982 to 0.928, although it was higher than the corresponding VGG16 model. While the prediction percentage of 93.1% was the lowest for DenseNet121 extractors, its precision of 0.955 and recall of 0.789 were higher than the VGG16 counterpart. The F1-score of 0.862 and AUC of 0.976 were also higher than the VGG16 model, indicating better overall performance.

For the Support Vector Machine (SVM) classifier with DenseNet121, accuracy dipped from 0.992 to 0.928, surpassing the VGG16 counterpart. Precision plunged from 0.992 to 0.916, although it remained higher than VGG16. Recall fell from 0.966 to 0.790, lower than the VGG16 model. However, the F1-score of 0.857 and AUC of 0.990 remained higher than the VGG16 model, suggesting better overall performance. The Random Forest (RF) classifier with DenseNet121 experienced a decrease in accuracy from 0.973 to 0.919, lower than the VGG16 model and the lowest among DenseNet121 extractors. Its prediction percentage tied for the lowest in DenseNet121 at 93.1%. While precision, recall, F1-score, and AUC were lower than the DenseNet121 Logistic Regression model, they were higher than the corresponding VGG16 model, indicating better generalization performance.

These observations further support the conclusions drawn in previous studies, emphasizing the impact of domain shift on the performance of machine learning models, particularly those based on deep learning, when applied to new and unseen datasets. However, it is worth noting that the DenseNet121 models generally outperformed their VGG16 counterparts, suggesting that the DenseNet121 feature extractor exhibits greater resilience to domain shift than VGG16. Among the classifiers considered, the Logistic Regression (LR) classifier with the DenseNet121 feature extractor consistently achieved the highest metrics across various categories, including accuracy, prediction percentage, precision, recall, F1-score, and AUC. This finding strengthens the observation that the combination of deep learning-based feature extraction and logistic regression classification may offer enhanced robustness to domain shift compared to other classifiers such as Random Forest, Gradient Boosting, or Support Vector Machines. To mitigate the effects of domain shift and improve the generalization capabilities of the hybrid models for malignant prediction, several strategies can be employed. These include data augmentation, transfer learning, domain adaptation techniques, and the incorporation of more diverse and representative data during training. Additionally, exploring alternative feature extractors, such as DenseNet121, which

demonstrate superior resistance to domain shift, could prove beneficial. By addressing domain shift and leveraging the most robust hybrid model architectures, the performance of malignant prediction can be further optimized. This ensures more consistent and reliable predictions for both benign and malignant cases when applying these models to new and unseen mammogram images in clinical practice.

4.2.4 Comparison with previous studies

In this section, we will conduct a comprehensive comparison between the performance of our best hybrid model, obtained in this study, and other state-of-the-art hybrid models reported in previous studies. This comparison is vital to validate the effectiveness of our proposed approach and to assess the potential improvements achieved in the field of breast cancer diagnosis using mammogram images. The previous studies we will consider for benchmarking are listed in Table 10. By evaluating our top-performing model against these established baselines, our objective is to determine whether our methodology has indeed achieved notable advancements in terms of diagnostic accuracy, precision, recall, F1-score, area under the curve (AUC), and other pertinent performance metrics. This comparative analysis will allow us to position our work within the broader context of existing research endeavors and quantify the extent to which our hybrid model surpasses or aligns with the current state-of-the-art techniques. Ultimately, this rigorous benchmarking process will provide valuable insights into the strengths and potential limitations of our approach, guiding future research directions, and facilitating the practical implementation of our model in clinical settings to enhance breast cancer screening and diagnosis.

Table 10: Other state-of-the-art hybrid models reported in previous.

Reference	Feature Extractor	Classifier	Results	
			ACC [%]	AUC
RUNYU SONG (2020)	Deep Convolutional Neural Network (DCNN)	Support Vector Machine (SVM) Gradient Boosting (XGBoost)	92.80%	-
Dina A. Ragab (2019)	AlexNet	Support Vector Machine (SVM)	71.01%	0.88
Dina A. Ragab (2021)	Deep Convolutional Neural Network (DCNN)	Support Vector Machine (SVM)	97.9%	-

All the studies included in this comparison aimed to develop advanced computer-aided diagnosis (CAD) systems for the detection and classification of breast cancer from mammogram images, a critical task in improving patient outcomes and survival rates. However, the methodologies and techniques employed in these studies exhibited significant variations. In our study, we adopted a cutting-edge approach by exploring hybrid artificial intelligence models that synergistically combined deep learning and machine learning techniques. Specifically, we leveraged the powerful feature extraction capabilities of pre-trained convolutional neural networks (CNNs), namely VGG16 and DenseNet121, to extract discriminative features from the mammogram images. These extracted features were then utilized to train various machine learning classifiers, including logistic regression, support vector machines (SVMs), random forests, and gradient boosting models. In contrast, the study conducted by Ardalan Ghasemzadeh (2018) employed a more traditional approach, relying on handcrafted feature extraction based on the Gabor wavelet transform to derive feature vectors from the mammogram images. These feature vectors were subsequently classified using a range of machine learning techniques, such as C5.0 decision tree, SVM, artificial neural networks, quest tree, and CHAID algorithm.

One of the key strengths of this study was its comprehensive evaluation framework, which involved testing the hybrid AI models across three distinct stages: (1) using the original mammogram images, (2) utilizing enhanced mammogram images processed with various image preprocessing and enhancement techniques, and (3) testing on new mammogram images. The image enhancement techniques employed included morphological erosion preprocessing, Contrast-Limited Adaptive Histogram Equalization (CLAHE), Laplacian of Gaussian (LoG) edge enhancement, and unsharp masking. This rigorous evaluation approach enabled a thorough assessment of the impact of image preprocessing and enhancement on the predictive capabilities of the hybrid models, a crucial aspect that was not explicitly explored in the other study.

The results obtained in this study were truly remarkable. When predicting benign cases from the enhanced mammogram images in stage 2, the logistic regression classifier with DenseNet121 features achieved outstanding performance, with the highest accuracy of 0.991, precision of 0.996, F1-score of 0.989, and AUC of 0.999. The SVM with DenseNet121 features also demonstrated exceptional performance, with an accuracy of 0.986 and an AUC of 0.999. Impressively, the logistic regression model with VGG16 features exhibited the shortest predictive time of 0.13

seconds, making it suitable for time-sensitive clinical applications. In the prediction of malignant cases from the enhanced images, the logistic regression classifier with DenseNet121 features once again excelled, achieving the highest accuracy of 0.995, precision of 0.995, recall of 0.995, F1-score of 0.995, and an AUC of 0.999. The SVM with DenseNet121 features closely followed with an accuracy of 0.992 and an AUC of 0.998. Notably, the logistic regression model with VGG16 features demonstrated the fastest predictive time of 0.08 seconds. In comparison, the best-performing classifier in the Ardalan Ghasemzadeh study was the SVM, which achieved an accuracy of 0.968, sensitivity of 0.997, and specificity of 0.94. While these results are respectable, they fall short of the exceptional performance achieved by the hybrid models in this study.

On the contrary, in the study conducted by RUNYU SONG (2020), a combined feature CAD method based on deep learning was proposed for the classification of mammographic masses into three classes: normal, benign, and malignant (cancer). This study utilized a Deep Convolutional Neural Network (DCNN) as a feature extractor to score the three types of breast masses. The scoring features from the DCNN were then combined with image texture features, including Gray-Level Co-occurrence Matrix (GLCM) and Histogram of Oriented Gradient (HOG), as inputs for the classification stage. The performance of two classifiers, Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost), was evaluated in terms of metrics such as accuracy (ACC), precision (Pre), recall (Rec), F1-score (F1), and overall accuracy (Overall ACC). The study by RUNYU SONG (2020) achieved its best results with the XGBoost classifier, obtaining an overall accuracy of 92.80% and a malignant tumor identification rate of 84%. While these results are respectable, they fall short of the exceptional performance achieved by the hybrid models in this study, particularly in terms of accuracy, precision, recall, F1-score, and AUC values for malignant case prediction.

Lastly, the studies conducted by Dina A. Ragab (2019 and 2021) proposed CAD systems primarily based on deep learning techniques for the classification of benign and malignant mass tumors in breast mammography images. The 2019 study employed a deep convolutional neural network (DCNN) architecture, specifically AlexNet, which was fine-tuned for binary classification. The last fully connected layer of AlexNet was connected to an SVM classifier in an attempt to improve accuracy. Two segmentation approaches were explored: manual region of interest (ROI) determination and threshold and region-based segmentation. Data augmentation techniques, such

as rotation, were also applied to increase the size of the input data. The accuracy of the fine-tuned DCNN architecture was 71.01% when cropping the ROI manually from the mammogram, with an AUC of 0.88 (88%) for samples obtained from both segmentation techniques.

In the 2021 study conducted by Dina A. Ragab, four experiments were carried out to determine the optimal approach. The first experiment involved using end-to-end pre-trained fine-tuned DCNN networks. In the second experiment, deep features were extracted from the DCNNs and fed into an SVM classifier with different kernel functions. The third experiment focused on deep feature fusion to demonstrate the enhanced accuracy achievable by combining deep features with SVM classifiers. Lastly, the fourth experiment introduced principal component analysis (PCA) to reduce the large feature vector resulting from feature fusion and decrease computational costs. According to the study by Dina A. Ragab (2021), deep feature fusion yielded the highest accuracy compared to state-of-the-art CAD systems for both datasets. However, applying PCA to the feature fusion sets did not improve accuracy, although it reduced computational costs by decreasing execution time. While the studies conducted by Dina A. Ragab (2019 and 2021) contributed to the development of CAD systems for breast cancer diagnosis, they did not achieve the exceptional performance demonstrated by the hybrid models in our study, particularly in terms of accuracy, precision, recall, F1-score, and AUC values for malignant case prediction. The comprehensive evaluation framework employed in our study, which involved testing across different stages and utilizing various image preprocessing and enhancement techniques, distinguishes it and highlights its potential for advancing the field of medical image analysis and computer-aided diagnosis for breast cancer.

Overall, this study showcases the potential of hybrid artificial intelligence approaches that effectively combine the strengths of deep learning for feature extraction and machine learning for classification. By achieving high accuracy, precision, recall, F1-scores, and AUC values in predicting breast cancer malignancy from mammogram images, your work holds significant promise for improving computer-aided diagnosis and decision support systems in breast cancer screening and diagnosis. The effective fusion of these techniques, along with the comprehensive evaluation framework and exploration of image preprocessing and enhancement techniques, sets your study apart and underscores its potential for advancing the field of medical image analysis and computer-aided diagnosis for breast cancer.

Chapter Six: Conclusion

The objective of this study was to investigate the efficacy of hybrid artificial intelligence models that combine deep learning and machine learning for the prediction of benign and malignant breast cancer using mammogram images. We employed VGG16 and DenseNet121, two deep learning models, to extract features from the mammogram images. Subsequently, we utilized machine learning algorithms, namely Random Forest, Gradient Boosting, Support Vector Machine, and Logistic Regression, to classify the images as either benign or malignant based on the extracted features. The performance of each hybrid model was evaluated using metrics such as accuracy, precision, recall, F1-score, AUC, and runtime.

Upon analyzing the results obtained from the original mammogram images (prior to applying image enhancement techniques), we observed promising outcomes for the hybrid AI models in predicting both benign and malignant breast cancer. In the prediction of benign cases, the DenseNet121 feature extractor generally outperformed the VGG16 feature extractor in terms of accuracy, precision, recall, F1-score, and AUC across all classifiers. Specifically, the DenseNet121 feature extractor demonstrated higher accuracy, precision, recall, and F1-score, indicating its superior ability to accurately classify benign cases. Among all the classifiers, the Gradient Boosting classifier with the DenseNet121 feature extractor achieved the highest accuracy (0.869) and AUC (0.858), highlighting its potential for accurate classification of benign cases. However, it takes a long time to predict, reaching up to 187.15 seconds.

In contrast, the performance of the hybrid models varied when predicting malignant cases. While the DenseNet121 feature extractor generally yielded higher accuracy, precision, recall, F1-score, and AUC compared to the VGG16 feature extractor, the performance of each classifier differed. The Logistic Regression classifier with the DenseNet121 feature extractor attained the highest accuracy (0.872) and Recall (0.564), demonstrating its effectiveness in classifying malignant cases. The prediction time is very short, taking as little as 0.1 second, making it highly applicable in the clinical field compared to other models that require more time.

After applying image enhancement techniques to enhance the quality of mammogram images, notable improvements were observed in the performance of the hybrid AI models. The enhanced images resulted in higher accuracy, precision, recall, and F1-score, indicating enhanced

classification results for both benign and malignant cases. Consistently, the DenseNet121 feature extractor outperformed the VGG16 feature extractor across all classifiers in terms of accuracy, precision, recall, F1-score, and AUC.

For the prediction of benign cases from the enhanced images, the Logistic Regression classifier with the DenseNet121 feature extractor achieved the highest accuracy (0.991) and F1-score (0.989), demonstrating its superior ability to accurately classify benign cases. Similarly, in the prediction of malignant cases, the Logistic Regression classifier with the DenseNet121 feature extractor attained the highest accuracy (0.995) and F1-score (0.995), underscoring its potential for accurate classification of malignant cases. In both scenarios, the prediction time was very short, taking as little as 0.15 seconds for benign breast cancer prediction and 0.22 seconds for malignant breast cancer prediction.

Furthermore, the proposed hybrid AI models were applied to a new dataset of mammogram images, which consisted of 400 benign and 400 malignant cases confirmed by biopsy results. The results indicated promising performance of the hybrid models in predicting both benign and malignant breast cancer. Across all classifiers, the DenseNet121 feature extractor exhibited superior performance in terms of accuracy, precision, recall, F1-score, and AUC compared to the VGG16 feature extractor. Specifically, the Logistic Regression classifier with the DenseNet121 feature extractor achieved the highest accuracy (0.967), AUC (0.999) and TIME (0.06 seconds) for classifying benign cases, and the highest accuracy (0.955), AUC (0.998) and TIME (0.17 seconds) for classifying malignant cases.

In summary, our study highlights the potential of hybrid artificial intelligence models that combine deep learning and machine learning for accurate prediction of benign and malignant breast cancer using mammogram images. The incorporation of image enhancement techniques further enhances the performance of these models, underscoring their effectiveness in real-world applications. However, it is important to acknowledge that the performance of the models may vary depending on the dataset and specific image characteristics. To validate these findings, future research should encompass larger and more diverse datasets, as well as explore the potential of alternative deep learning and machine learning techniques for breast cancer prediction. Overall, the outcomes of this study contribute to the advancement of AI-driven approaches in breast cancer diagnosis and hold implications for improving early detection and treatment outcomes.

6.1 Strength and Limitations

The study conducted extensive research on the use of hybrid artificial intelligence (AI) models for predicting benign and malignant breast cancer from mammogram images. By combining deep learning and machine learning techniques, the study introduced a novel approach that aimed to leverage the strengths of both methods. The incorporation of image enhancement techniques, such as CLAHE, LoG edge enhancement, and unsharp masking, significantly improved the visibility of important structures and features in mammogram images. Consequently, the predictive capabilities of the AI models were enhanced. The study reported high-performance metrics, indicating the potential of the proposed approach in achieving accurate and reliable predictions of breast cancer malignancy. Furthermore, certain models demonstrated efficient prediction times, making them suitable for real-time clinical applications.

However, there were some limitations to consider. The dataset used for training and testing the hybrid AI models may not have been sufficiently large or diverse, which could affect the models' performance. The generalizability of the models to diverse patient populations, imaging modalities, and clinical settings was not fully explored. The study also did not address the interpretability and explain-ability of the models' decision-making processes, which is crucial for building trust and facilitating clinical adoption. Additionally, potential biases in the dataset or models were not explicitly addressed. Moreover, the study did not provide details on the practical aspects of clinical integration and implementation.

6.2 Recommendations

The researchers recommend acquiring a larger and more diverse mammogram image dataset to enhance the generalizability and robustness of the hybrid AI models. Conducting external validation studies using independent datasets from multiple healthcare institutions would provide a more realistic assessment of the models' performance and transferability. Comparative analysis with existing state-of-the-art approaches and radiologists' assessments is also recommended. To enhance model interpretability and explain-ability, incorporating techniques from the field of explainable AI (XAI) is advised. Optimizing computational resources and exploring strategies like model compression, pruning, or hardware acceleration would ensure efficient training and deployment. Expanding the scope to incorporate multimodal data, including ultrasound, MRI, and patient history, could potentially enhance predictive power and clinical utility.

It is highly recommended to incorporate advanced image enhancement techniques when developing hybrid AI models for breast cancer detection and classification. Close collaboration with radiologists and breast imaging specialists is crucial to ensure interpretability and alignment with clinical decision-making processes. Mitigating potential overfitting issues through techniques like regularization, data augmentation, or transfer learning approaches is essential. Addressing ethical and regulatory considerations such as data privacy, bias mitigation, transparency, and liability is also necessary.

6.3 Future study

Future research can explore conducting multicenter studies involving a larger and more diverse dataset to provide a comprehensive evaluation of the hybrid AI models' performance and generalizability. Extending the study by integrating clinical data and multimodal imaging data could provide a more holistic representation of breast cancer and enhance accuracy and reliability. Investigating the application of the hybrid AI models for longitudinal analysis and monitoring of breast cancer progression or regression is another avenue. Exploring advanced techniques from explainable AI to improve the interpretability and transparency of the deep learning models used for feature extraction is recommended.

Conducting prospective clinical validation studies to evaluate the models' performance in real-world clinical settings and integrating them into existing workflows is crucial. Performing cost-effectiveness and health economic analyses to evaluate the potential economic impact of implementing the hybrid AI models is also recommended. Investigating the application of advanced machine learning techniques, such as attention mechanisms, adversarial training, or self-supervised learning, can further enhance the models' performance and robustness. Fostering collaborations with researchers, clinicians, and industry partners to share knowledge, datasets, and best practices is essential. Exploring the integration of the hybrid AI models into comprehensive decision support systems for breast cancer diagnosis and treatment planning is recommended. Additionally, future studies could explore integrating automated lesion segmentation capabilities into the hybrid AI models to provide a comprehensive system for prediction and localization of lesions.

References

A. M. S. C. Meenakshi, M. Govindarajan, “Detection of breast cancer using MLP and RBF classifiers,” IMS Manthan, vol. 5, no. 1, 2010.

Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial intelligence in Medicine*, *25*(3), 265-281.

ABBOODI, C. H. (2014). MAMMOGRAM IMAGE ENHANCEMENT BY USING A TWO-STAGE DENOISING FILTER AND CONTRAST LIMITED ADAPTIVE HISTOGRAM EQUALIZATION.

ABD ALMALEKI, P., Yarmohammadi, M., & GITI, M. (2004). Comparison of logistic regression and neural network models in predicting the outcome of biopsy in breast cancer from MRI findings.

Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., & Abdel-Mottaleb, M. (2021). Convolutional neural networks for breast cancer detection in mammography: A survey. *Computers in biology and medicine*, *131*, 104248.

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. *Expert Systems with Applications*, *46*, 139-144.

Abdulloh, A., & Ni'mah, A. Q. (2023). BI-RADS CLASSIFICATION FOR BREAST ULTRASOUND: A REVIEW. *PHARMACOLOGY, MEDICAL REPORTS, ORTHOPEDIC, AND ILLNESS DETAILS (COMORBID)*, *2*(2), 67-84.

ACR, A. C. o. R. (2023). *ACR BI-RADS® Atlas Fifth Edition*. Retrieved 16/04/2024 from https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS-Poster.pdf.

ACS. (2021). *About Breast Cancer*. American Cancer Society. Retrieved 17 November 2023 from https://www.cancer.org/content/dam/CRC/PDF/Public/8577.00.pdf.

Agrawal, S., Rangnekar, R., Gala, D., Paul, S., & Kalbande, D. (2018). Detection of breast cancer from mammograms using a hybrid approach of deep learning and linear classification. 2018 International Conference on Smart City and Emerging Technology (ICSCET).

Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, *7*, e7702.

Al-Antari, M. A., Al-Masni, M. A., Choi, M.-T., Han, S.-M., & Kim, T.-S. (2018). A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *International journal of medical informatics*, *117*, 44-54.

Al-Antari, M. A., Han, S.-M., & Kim, T.-S. (2020). Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. *Computer methods and programs in biomedicine*, *196*, 105584.

Alarabeyyat, A., & Alhanahnah, M. (2016). Breast cancer detection using k-nearest neighbor machine learning algorithm. 2016 9th International Conference on Developments in eSystems Engineering (DeSE).

Al-Masni, M. A., Al-Antari, M. A., Park, J.-M., Gi, G., Kim, T.-Y., Rivera, P., Valarezo, E., Choi, M.-T., Han, S.-M., & Kim, T.-S. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer methods and programs in biomedicine*, *157*, 85-94.

Al-Masni, M. A., Al-Antari, M. A., Park, J.-M., Gi, G., Kim, T.-Y., Rivera, P., Valarezo, E., Choi, M.-T., Han, S.-M., & Kim, T.-S. (2018). Simultaneous detection and classification of

breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer methods and programs in biomedicine*, *157*, 85-94.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.

Alonso, A., & Siracuse, J. J. (2023). Protecting patient safety and privacy in the era of artificial intelligence. Seminars in Vascular Surgery.

Alshayeji, M. H., Ellethy, H., & Gupta, R. (2022). Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach. *Biomedical Signal Processing and Control*, *71*, 103141.

Aly, G. H., Marey, M., El-Sayed, S. A., & Tolba, M. F. (2021). YOLO based breast masses detection and classification in full-field digital mammograms. *Computer methods and programs in biomedicine*, *200*, 105823.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, *8*, 1-74.

Anand, R., Shanthi, T., Nithish, M., & Lakshman, S. (2020). Face recognition and classification using GoogleNET architecture. Soft Computing for Problem Solving: SocProS 2018, Volume 1.

Anandhamala, G. (2018). Recent trends in medical imaging modalities and challenges for diagnosing breast cancer. *Biomedical and Pharmacology Journal*, *11*(3), 1649-1658.

Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2015). Convolutional neural networks for mammography mass lesion classification. 2015 37th Annual international conference of the IEEE engineering in medicine and biology society (EMBC).

Arora, R., Rai, P. K., & Raman, B. (2020). Deep feature–based automatic classification of mammograms. *Medical & biological engineering & computing*, *58*, 1199-1211.

Arora, R., Rai, P. K., & Raman, B. (2020). Deep feature–based automatic classification of mammograms. *Medical & biological engineering & computing*, *58*, 1199-1211.

Ayer, T., Chhatwal, J., Alagoz, O., Kahn Jr, C. E., Woods, R. W., & Burnside, E. S. (2010). Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *Radiographics*, *30*(1), 13-22.

Azar, A. T., & El-Said, S. A. (2013). Probabilistic neural network for breast cancer classification. *Neural Computing and Applications*, *23*, 1737-1751.

Azar, A. T., & El-Said, S. A. (2014). Performance analysis of support vector machines classifiers in breast cancer mammography recognition. *Neural Computing and Applications*, *24*, 1163-1177.

Azhdeh, S., Kaviani, A., Sadighi, N., & Rahmani, M. (2021). Accurate estimation of breast tumor size: a comparison between ultrasonography, mammography, magnetic resonance imaging, and associated contributing factors. *European Journal of Breast Health*, *17*(1), 53.

B. Dai, R.-C. Chen, S.-Z. Zhu, and W.-W. Zhang, "Using random forest algorithm for breast cancer diagnosis," in 2018 International Symposium on Computer, Consumer and Control (IS3C), Dec. 2018, pp. 449–452, doi: 10.1109/IS3C.2018.00119.

Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., & Rudin, C. (2021). A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, *3*(12), 1061-1070.

Basha, S. S., Dubey, S. R., Pulabaigari, V., & Mukherjee, S. (2020). Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, *378*, 112-119.

Beauxis-Aussalet, E., & Hardman, L. (2014). Simplifying the visualization of confusion matrix. 26th Benelux conference on artificial intelligence (BNAIC).

Becker, A. S., Marcon, M., Ghafoor, S., Wurnig, M. C., Frauenfelder, T., & Boss, A. (2017). Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investigative radiology*, *52*(7), 434-440.

Bektaş, B., Emre, İ. E., Kartal, E., & Gulsecen, S. (2018). Classification of mammography images by machine learning techniques. 2018 3rd International Conference on Computer Science and Engineering (UBMK).

Berbar, M. A. (2018). Hybrid methods for feature extraction for breast masses classification. *Egyptian informatics journal*, *19*(1), 63-73.

Boden, M. A. (1996). *Artificial intelligence*. Elsevier.

Bokade, A., & Shah, A. (2021). Breast mass classification with deep transfer feature extractor model and random forest classifier. 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT).

C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," Journal of Biomedical Science and Engineering, vol. 6, no. 5, pp. 551–560, 2013, doi: 10.4236/jbise.2013.65070.

Cao, J., Cui, H., Zhang, Q., & Zhang, Z. (2020). Ancient mural classification method based on improved AlexNet network. *Studies in Conservation*, *65*(7), 411-423.

Chakraborty, J., Midya, A., & Rabidas, R. (2018). Computer-aided detection and diagnosis of mammographic masses using multi-resolution analysis of oriented tissue patterns. *Expert Systems with Applications*, *99*, 168-179.

Charan, S., Khan, M. J., & Khurshid, K. (2018). Breast cancer detection in mammograms using convolutional neural network. 2018 international conference on computing, mathematics and engineering technologies (iCoMET).

Chebli, A., Djebbar, A., & Marouani, H. F. (2018). Semi-supervised learning for medical application: A survey. 2018 international conference on applied smart systems (ICASS).

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1-13.

Choi, J. Y., Kim, D. H., Plataniotis, K. N., & Ro, Y. M. (2016). Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography. *Expert Systems with Applications*, *46*, 106-121.

Chu, J., Min, H., Liu, L., & Lu, W. (2015). A novel computer aided breast mass detection scheme based on morphological enhancement and SLIC superpixel segmentation. *Medical physics*, *42*(7), 3859-3869.

Coppola, F., Faggioni, L., Gabelloni, M., De Vietro, F., Mendola, V., Cattabriga, A., Cocozza, M. A., Vara, G., Piccinino, A., & Lo Monaco, S. (2021). Human, all too human? An all-around appraisal of the "artificial intelligence revolution" in medical imaging. *Frontiers in psychology*, *12*, 710982.

Craciunescu, O. I., Blackwell, K. L., Jones, E. L., MacFall, J. R., Yu, D., Vujaskovic, Z., Wong, T. Z., Liotcheva, V., Rosen, E. L., & Prosnitz, L. R. (2009). DCE-MRI parameters have potential to predict response of locally advanced breast cancer patients to neoadjuvant chemotherapy and hyperthermia: a pilot study. *International Journal of Hyperthermia*, *25*(6), 405-415.

D. Soria, J. M. Garibaldi, E. Biganzoli, and I. O. Ellis, "A comparison of three different methods for classification of breast cancer data," in 2008 Seventh International Conference on Machine Learning and Applications, 2008, pp. 619–624, doi: 10.1109/ICMLA.2008.97.

da Rocha, S. V., Junior, G. B., Silva, A. C., de Paiva, A. C., & Gattass, M. (2016). Texture analysis of masses malignant in mammograms images using a combined approach of diversity index and local binary patterns distribution. *Expert Systems with Applications*, *66*, 7-19.

Danala, G., Aghaei, F., Heidari, M., Wu, T., Patel, B., & Zheng, B. (2018). Computer-aided classification of breast masses using contrast-enhanced digital mammograms. Medical Imaging 2018: Computer-Aided Diagnosis.

de Brito Silva, T. F., de Paiva, A. C., Silva, A. C., Braz Júnior, G., & de Almeida, J. D. S. (2020). Classification of breast masses in mammograms using geometric and topological feature maps and shape distribution. *Research on Biomedical Engineering*, *36*, 225-235.

de Nazaré Silva, J., de Carvalho Filho, A. O., Corrêa Silva, A., Cardoso de Paiva, A., & Gattass, M. (2015). Automatic detection of masses in mammograms using quality threshold clustering, correlogram function, and SVM. *Journal of digital imaging*, *28*, 323-337.

de Sampaio, W. B., Silva, A. C., de Paiva, A. C., & Gattass, M. (2015). Detection of masses in mammograms with adaption to breast density using genetic algorithm, phylogenetic trees, LBP and SVM. *Expert Systems with Applications*, *42*(22), 8911-8928.

DeMartini, W. B., Kurland, B. F., Gutierrez, R. L., Blackmore, C. C., Peacock, S., & Lehman, C. D. (2011). Probability of malignancy for lesions detected on breast MRI: a predictive model incorporating BI-RADS imaging features and patient characteristics. *European radiology*, *21*, 1609-1617.

Dhahbi, S., Barhoumi, W., & Zagrouba, E. (2015). Breast cancer diagnosis in digitized mammograms using curvelet moments. *Computers in biology and medicine*, *64*, 79-90.

Dhahbi, S., Barhoumi, W., Kurek, J., Swiderski, B., Kruk, M., & Zagrouba, E. (2018). False-positive reduction in computer-aided mass detection using mammographic texture analysis and classification. *Computer methods and programs in biomedicine*, *160*, 75-83.

Dheeba, J., Singh, N. A., & Selvi, S. T. (2014). Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *Journal of biomedical informatics*, *49*, 45-52.

Dhungel, N., Carneiro, G., & Bradley, A. P. (2015). Automated mass detection in mammograms using cascaded deep learning and random forests. 2015 international conference on digital image computing: techniques and applications (DICTA).

Dhungel, N., Carneiro, G., & Bradley, A. P. (2017). A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical image analysis*, *37*, 114-128.

Duan, X., Mei, Y., Wu, S., Ling, Q., Qin, G., Ma, J., Chen, C., Qi, H., Zhou, L., & Xu, Y. (2018). A multiscale contrast enhancement for mammogram using dynamic unsharp masking in Laplacian pyramid. *IEEE transactions on radiation and plasma medical sciences*, *3*(5), 557-564.

Duran-Lopez, L., Dominguez-Morales, J. P., Conde-Martin, A. F., Vicente-Diaz, S., & Linares-Barranco, A. (2020). PROMETEO: A CNN-based computer-aided diagnosis system for WSI prostate cancer detection. *IEEE Access*, *8*, 128613-128628.

Ekpo, E. U., Alakhras, M., & Brennan, P. (2018). Errors in mammography cannot be solved through technology alone. *Asian Pacific journal of cancer prevention: APJCP*, *19*(2), 291.

El_Rahman, S. A. (2021). Predicting breast cancer survivability based on machine learning and features selection algorithms: a comparative study. *Journal of Ambient Intelligence and Humanized Computing*, *12*(8), 8585-8623.

Eltoukhy, M. M., Elhoseny, M., Hosny, K. M., & Singh, A. K. (2018). Computer aided detection of mammographic mass using exact Gaussian–Hermite moments. *Journal of Ambient Intelligence and Humanized Computing*, 1-9.

Francies, F. Z., Hull, R., Khanyile, R., & Dlamini, Z. (2020). Breast cancer in low-middle income countries: abnormality in splicing and lack of targeted treatment options. *American journal of cancer research*, *10*(5), 1568.

Gao, F., Wu, T., Li, J., Zheng, B., Ruan, L., Shang, D., & Patel, B. (2018). SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis. *Computerized Medical Imaging and Graphics*, *70*, 53-62.

Gao, H., Zhen, T., & Li, Z. (2022). Detection of wheat unsound kernels based on improved ResNet. *IEEE Access*, *10*, 20092-20101.

Gardezi, S. J. S., Elazab, A., Lei, B., & Wang, T. (2019). Breast cancer detection and diagnosis using mammographic data: Systematic review. *Journal of medical Internet research*, *21*(7), e14464.

Gayathri, B., Sumathi, C., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms-a survey. *International Journal of Distributed and Parallel Systems*, *4*(3), 105.

Ghiasi, M. M., & Zendehboudi, S. (2021). Application of decision tree-based ensemble learning in the classification of breast cancer. *Computers in biology and medicine*, *128*, 104089.

Giampietro, R. R., Cabral, M. V. G., Lima, S. A. M., Weber, S. A. T., & dos Santos Nunes-Nogueira, V. (2020). Accuracy and effectiveness of mammography versus mammography and tomosynthesis for population-based breast cancer screening: a systematic review and meta-analysis. *Scientific reports*, *10*(1), 7991.

Gnanasekaran, V. S., Joypaul, S., Meenakshi Sundaram, P., & Chairman, D. D. (2020). Deep learning algorithm for breast masses classification in mammograms. *IET Image Processing*, *14*(12), 2860-2868.

Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, *56*(11), 12561-12605.

Gonçalves, C. B., Souza, J. R., & Fernandes, H. (2021). Classification of static infrared images using pre-trained CNN for breast cancer detection. 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS).

Günes, M. E. (2018). Comparison of the ultrasound-guided tru-cut biopsy with postoperative histopathology results in patients with breast mass. *Annali italiani di chirurgia*, *89*, 30-35.

Gupta, K., & Chawla, N. (2020). Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained CNN. *Procedia Computer Science*, *167*, 878-889.

Hah, H., & Goldin, D. S. (2021). How clinicians perceive artificial intelligence–assisted technologies in diagnostic decision making: Mixed methods approach. *Journal of Medical Internet Research*, *23*(12), e33540.

Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging*, *29*(3), 307-335.

He, H. (2020). A deep research in classic classification network. IOP Conference Series: Materials Science and Engineering.

Healthineers, S. (2023). *Mammomat Revelation*. Siemens Healthineers. Retrieved 20/01/2024 from https://www.siemens-healthineers.com/mammography/digital-mammography/mammomat-revelation.

Heinig, J., Witteler, R., Schmitz, R., Kiesel, L., & Steinhard, J. (2008). Accuracy of classification of breast ultrasound findings based on criteria used for BI-RADS. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, *32*(4), 573-578.

Heywang-Köbrunner, S. H., Hacker, A., & Sedlacek, S. (2011). Advantages and disadvantages of mammography screening. *Breast care*, *6*(3), 199-207.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, *12*(1), 5979.

Hiran, K. K., Jain, R. K., Lakhwani, K., & Doshi, R. (2021). *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*. BPB Publications.

Hlávka, J. P. (2020). Security, privacy, and information-sharing aspects of healthcare artificial intelligence. In *Artificial intelligence in healthcare* (pp. 235-270). Elsevier.

Horton, S., Camacho Rodriguez, R., Anderson, B. O., Aung, S., Awuah, B., Delgado Pebe, L., Duggan, C., Dvaladze, A., Kumar, S., & Murillo, R. (2020). Health system strengthening: Integration of breast cancer care for improved outcomes. *Cancer*, *126*, 2353-2364.

Houssein, E. H., Emam, M. M., & Ali, A. A. (2022). An optimized deep learning architecture for breast cancer diagnosis based on improved marine predators algorithm. *Neural computing and applications*, *34*(20), 18015-18033.

Hughes, C., & Hughes, T. (2019). What metrics should we use to measure commercial AI? *AI Matters*, *5*(2), 41-45.

Intriago-Pazmiño, M., Ibarra-Fiallo, J., Guzmán-Castillo, A., Alonso-Calvo, R., & Crespo, J. (2023). Quantitative Measures for Medical Fundus and Mammography Images Enhancement.

Ito, Y., Miyoshi, A., Ueda, Y., Tanaka, Y., Nakae, R., Morimoto, A., Shiomi, M., Enomoto, T., Sekine, M., & Sasagawa, T. (2022). An artificial intelligence-assisted diagnostic system improves the accuracy of image diagnosis of uterine cervical lesions. *Molecular and Clinical Oncology*, *16*(2), 1-6.

J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan, "Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis," IEEE Access, vol. 8, pp. 96946–96954, 2020, doi: 10.1109/ACCESS.2020.2993536.

Jafari, Z., & Karami, E. (2023). Breast cancer detection in mammography images: A CNN-based approach with feature selection. *Information*, *14*(7), 410.

Jalalian, A., Mashohor, S. B., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., & Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical imaging*, *37*(3), 420-426.

Jiao, Z., Gao, X., Wang, Y., & Li, J. (2016). A deep feature based framework for breast masses classification. *Neurocomputing*, *197*, 221-231.

Jie, H. J., & Wanda, P. (2020). RunPool: A dynamic pooling layer for convolution neural network. *International Journal of Computational Intelligence Systems*, *13*(1), 66-76.

Jung, H., Kim, B., Lee, I., Yoo, M., Lee, J., Ham, S., Woo, O., & Kang, J. (2018). Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PloS one*, *13*(9), e0203355.

Justaniah, E., Aldabbagh, G., Alhothali, A., & Abourokbah, N. (2022). Classifying Breast Density from Mammogram with Pretrained CNNs and Weighted Average Ensembles. *Applied Sciences*, *12*(11), 5599.

Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019). Cancer classification using gaussian naive bayes algorithm. 2019 international engineering conference (IEC).

Karabatak, M. (2015). A new classifier for breast cancer detection based on Naïve Bayesian. *Measurement*, *72*, 32-36.

Karellas, A., & Vedantham, S. (2008). Breast cancer imaging: a perspective for the next decade. *Medical physics*, *35*(11), 4878-4897.

Kashyap, D., Pal, D., Sharma, R., Garg, V. K., Goel, N., Koundal, D., Zaguia, A., Koundal, S., & Belay, A. (2022). Global increase in breast cancer incidence: risk factors and preventive measures. *BioMed research international*, *2022*.

Khan, H. N., Shahid, A. R., Raza, B., Dar, A. H., & Alquhayz, H. (2019). Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access*, *7*, 165724-165733.

Kharel, N., Alsadoon, A., Prasad, P., & Elchouemi, A. (2017). Early diagnosis of breast cancer using contrast limited adaptive histogram equalization (CLAHE) and Morphology methods. 2017 8th International Conference on Information and Communication Systems (ICICS).

Khorshid, S. F., & Abdulazeez, A. M. (2021). Breast cancer diagnosis based on k-nearest neighbors: a review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *18*(4), 1927-1951.

Kuhl, C. K. (2015). The changing world of breast cancer: a radiologist's perspective. *Investigative radiology*, *50*(9), 615-628.

Kumar, M., Singhal, S., Shekhar, S., Sharma, B., & Srivastava, G. (2022). Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. *Sustainability*, *14*(21), 13998.

L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in 2018 International Conference on Robots and Intelligent System (ICRIS), May 2018, pp. 157–160, doi: 10.1109/ICRIS.2018.00049.

Le, E., Wang, Y., Huang, Y., Hickman, S., & Gilbert, F. (2019). Artificial intelligence in breast imaging. *Clinical radiology*, *74*(5), 357-366.

Lim, W. K., & Er, M. J. (2004). Classification of mammographic masses using generalized dynamic fuzzy neural networks. *Medical physics*, *31*(5), 1288-1295.

Liu, Z., Peng, J., Guo, X., Chen, S., & Liu, L. (2024). Breast cancer classification method based on improved VGG16 using mammography images. *Journal of Radiation Research and Applied Sciences*, *17*(2), 100885.

M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 36, no. 2, pp. 3240–3247, Mar. 2009, doi: 10.1016/j.eswa.2008.01.009.

M. Hosni, I. Abnane, A. Idri, J. M. Carrillo de Gea, and J. L. Fernández Alemán, "Reviewing ensemble classification methods in breast cancer," Computer Methods and Programs in Biomedicine, vol. 177, pp. 89–112, Aug. 2019, doi: 10.1016/j.cmpb.2019.05.019.

M. I. H. Showrov, M. T. Islam, M. D. Hossain, and M. S. Ahmed, "Performance comparison of three classifiers for the classification of breast cancer dataset," in 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Dec. 2019, pp. 1–5, doi: 10.1109/EICT48899.2019.9068816.

M. Karabatak, "A new classifier for breast cancer detection based on Naïve Bayesian," Measurement, vol. 72, pp. 32–36, Aug. 2015, doi: 10.1016/j.measurement.2015.04.028.

M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast cancer detection using k-nearest neighbor machine learning algorithm," in 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Aug. 2016, pp. 35–39, doi: 10.1109/DeSE.2016.8.

Maitra, I. K., Nag, S., & Bandyopadhyay, S. K. (2012). A novel edge detection algorithm for digital mammogram. *International Journal of Information and Communication Technology Research*, *2*(2).

Makandar, A., & Halalli, B. (2015). A review on preprocessing techniques for digital mammography images. *International Journal of Computer Applications*, *975*, 8887.

Meenalochini, G., & Ramkumar, S. (2021). Survey of machine learning algorithms for breast cancer detection using mammogram images. *Materials Today: Proceedings*, *37*, 2738-2743.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Minkowitz, S., Moskowitz, R., Khafif, R. A., & Alderete, M. N. (1986). Tru-cut needle biopsy of the breast. An analysis of its specificity and sensitivity. *Cancer*, *57*(2), 320-323.

Mohanty, F., Rup, S., Dash, B., Majhi, B., & Swamy, M. (2019). A computer-aided diagnosis system using Tchebichef features and improved grey wolf optimized extreme learning machine. *Applied Intelligence*, *49*, 983-1001.

Mokhtari, S., Yen, K. K., & Liu, J. (2021). Effectiveness of artificial intelligence in stock market prediction based on machine learning. *arXiv preprint arXiv:2107.01031*.

Muhammad, U., Wang, W., Chattha, S. P., & Ali, S. (2018). Pre-trained VGGNet architecture for remote-sensing image scene classification. 2018 24th International Conference on Pattern Recognition (ICPR).

Müller, P., & Braun, A. (2023). Local performance evaluation of AI-algorithms with the generalized spatial recall index. *tm-Technisches Messen*, *90*(7-8), 464-477.

Murtaza, G., Shuib, L., Abdul Wahab, A. W., Mujtaba, G., Mujtaba, G., Nweke, H. F., Al-garadi, M. A., Zulfiqar, F., Raza, G., & Azmi, N. A. (2020). Deep learning-based breast cancer

classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, *53*, 1655-1720.

N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer," Annals of Medicine and Surgery, vol. 62, pp. 53–64, Feb. 2021, doi: 10.1016/j.amsu.2020.12.043.

N. Arya and S. Saha, "Multi-modal classification for human breast cancer prognosis prediction: Proposal of deep-learning based stacked ensemble model," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, doi: 10.1109/TCBB.2020.3018467.

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. 2017 15th international conference on ICT and knowledge engineering (ICT&KE).

P. P. Sengar, M. J. Gaikwad, and A. S. Nagdive, "Comparative study of machine learning algorithms for breast cancer prediction," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Aug. 2020, pp. 796–801, doi: 10.1109/ICSSIT48917.2020.9214267.

Pacilè, S., Lopez, J., Chone, P., Bertinotti, T., Grouin, J. M., & Fillard, P. (2020). Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiology: Artificial Intelligence*, *2*(6), e190208.

Pattanaik, R. K., Mishra, S., Siddique, M., Gopikrishna, T., & Satapathy, S. (2022). Breast cancer classification from mammogram images using extreme learning machine-based DenseNet121 model. *Journal of Sensors*, *2022*.

PHIC, P. H. I. C. (2023). *Health Annual Report, Palestine 2022*. Palestine: Palestinian Ministry of Health Retrieved from https://site.moh.ps/.

Pinheiro, J. M. H., & Becker, M. (2024). Breast Cancer Classification Using Gradient Boosting Algorithms Focusing on Reducing the False Negative and SHAP for Explainability. *arXiv preprint arXiv:2403.09548*.

Pisano, E. D., Zong, S., Hemminger, B. M., DeLuca, M., Johnston, R. E., Muller, K., Braeuning, M. P., & Pizer, S. M. (1998). Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of digital imaging*, *11*, 193-200.

R. Nagarajan and M. Upreti, "An ensemble predictive modeling framework for breast cancer classification," Methods, vol. 131, pp. 128–134, Dec. 2017, doi: 10.1016/j.ymeth.2017.07.011.

Ragab, D. A., Attallah, O., Sharkas, M., Ren, J., & Marshall, S. (2021). A framework for breast cancer classification using multi-DCNNs. *Computers in biology and medicine*, *131*, 104245.

Ragab, D. A., Sharkas, M., Marshall, S., & Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ*, *7*, e6201.

Raza, K., & Singh, N. K. (2021). A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging*, *17*(9), 1059-1077.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, *8*(1), 4165.

Rubin, D. L. (2019). Artificial intelligence in imaging: the radiologist's role. *Journal of the American College of Radiology*, *16*(9), 1309-1317.

Rybiałek, A., & Jeleń, Ł. (2020). Application of densenets for classification of breast cancer mammograms. International Conference on Computer Information Systems and Industrial Management.

S. J. Malebary and A. Hashmi, "Automated breast mass classification system using deep learning and ensemble learning in digital mammogram," IEEE Access, vol. 9, pp. 55312–55328, 2021, doi: 10.1109/ACCESS.2021.3071297.

S. Kabiraj et al., "Breast cancer risk prediction using XGBoost and random forest algorithm," in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1–4, doi: 10.1109/ICCCNT49239.2020.9225451.

S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Dec. 2018, pp. 114–118, doi: 10.1109/CTEMS.2018.8769187.

S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," Applied Soft Computing, vol. 86, Jan. 2020, doi: 10.1016/j.asoc.2019.105941.

Sadoughi, F., Kazemy, Z., Hamedan, F., Owji, L., Rahmanikatigari, M., & Azadboni, T. T. (2018). Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. *Breast Cancer: Targets and Therapy*, 219-230.

Saint John's Cancer Institute. (2024). *Types of Breast Cancer*. Saint John's Cancer Institute. Retrieved 17 November 2023 from https://www.saintjohnscancer.org/breast/breast-cancer/types-of-breast-cancer/.

Sarker, I. H. (2021). Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. *SN Computer Science*, *2*(3), 154.

Sathiyanarayanan, P., Pavithra, S., Saranya, M. S., & M.akeswari, M. (2019). Identification of breast cancer using the decision tree algorithm. 2019 IEEE International conference on system, computation, automation and networking (ICSCAN).

Seryasat, O. R., & Haddadnia, J. (2018). Evaluation of a new ensemble learning framework for mass classification in mammograms. *Clinical breast cancer*, *18*(3), e407-e420.

Shah, S. M., Khan, R. A., Arif, S., & Sajid, U. (2022). Artificial intelligence for breast cancer analysis: Trends & directions. *Computers in biology and medicine*, *142*, 105221.

Shaik, S., & Kirthiga, S. (2021). Automatic modulation classification using DenseNet. 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP).

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, *9*(1), 12495.

Sheth, D., & Giger, M. L. (2020). Artificial intelligence in the interpretation of breast cancer on MRI. *Journal of Magnetic Resonance Imaging*, *51*(5), 1310-1324.

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: a cancer journal for clinicians*, *68*(1), 7-30.

Sivanantham, E., Epsiba, P., Gopi, B., Solainayagi, P., Umapathy, K., & Kumar, S. M. (2023). Mammogram classification using VGG-16 architecture. AIP Conference Proceedings.

Smithuis, H. Z. a. R. (2014). *Bi-RADS for Mammography and Ultrasound 2013*. Radiology Assistant. Retrieved 15/04/2024 from https://radiologyassistant.nl/breast/bi-rads/bi-rads-for-mammography-and-ultrasound-2013.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Australasian joint conference on artificial intelligence.

Soltani, R., Goeckel, D., Towsley, D., & Houmansadr, A. (2019). Fundamental limits of invisible flow fingerprinting. *IEEE Transactions on Information Forensics and Security*, *15*, 345-360.

Song, R., Li, T., & Wang, Y. (2020). Mammographic classification based on XGBoost and DCNN with multi features. *IEEE Access*, *8*, 75011-75021.

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, *9*, 1-10.

Strigel, R. M., Rollenhagen, J., Burnside, E. S., Elezaby, M., Fowler, A. M., Kelcz, F., Salkowski, L., & DeMartini, W. B. (2017). Screening breast MRI outcomes in routine clinical practice: comparison to BI-RADS benchmarks. *Academic radiology*, *24*(4), 411-417.

Strobel, K., Schrading, S., Hansen, N. L., Barabasch, A., & Kuhl, C. K. (2015). Assessment of BI-RADS category 4 lesions detected with screening mammography and screening US: utility of MR imaging. *Radiology*, *274*(2), 343-351.

Sun, M., Song, Z., Jiang, X., Pan, J., & Pang, Y. (2017). Learning pooling for convolutional neural network. *Neurocomputing*, *224*, 96-104.

Sun, Y., Xue, B., Zhang, M., & Yen, G. G. (2019). Completely automated CNN architecture design based on blocks. *IEEE transactions on neural networks and learning systems*, *31*(4), 1242-1254.

T. Price and N. Lindqvist, "Evaluation of feature selection methods for machine learning classification of breast cancer," Degree Project, Kth Royal Institute of Technology, School of Electrical Engineering and Computer Science. Stockholm, Sweden, 2018.

Tabrizchi, H., Tabrizchi, M., & Tabrizchi, H. (2020). Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree. *SN Applied Sciences*, *2*(4), 752.

Tahmasbi, A., Saki, F., & Shokouhi, S. B. (2011). Classification of benign and malignant masses based on Zernike moments. *Computers in biology and medicine*, *41*(8), 726-735.

Tai, S.-C., Chen, Z.-S., & Tsai, W.-T. (2013). An automatic mass detection system in mammograms based on complex texture features. *IEEE journal of biomedical and health informatics*, *18*(2), 618-627.

Tang, J., Rangayyan, R. M., Xu, J., El Naqa, I., & Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE transactions on information technology in biomedicine*, *13*(2), 236-251.

Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

TEAM, H. J. (2018). *Breast biopsy*. HEALTH JADE. Retrieved 17/09/2023 from https://healthjade.com/breast-biopsy/.

Thein, H. T. T., & Tun, K. M. M. (2015). An approach for breast cancer diagnosis classification using neural network. *Advanced Computing*, *6*(1), 1.

Tsopra, R., Fernandez, X., Luchinat, C., Alberghina, L., Lehrach, H., Vanoni, M., Dreher, F., Sezerman, O. U., Cuggia, M., & de Tayrac, M. (2021). A framework for validating AI in

precision medicine: considerations from the European ITFoC consortium. *BMC medical informatics and decision making*, *21*, 1-14.

V. Chaurasia and S. Pal, "Stacking-based ensemble framework and feature selection technique for the detection of breast cancer," SN Computer Science, vol. 2, no. 2, Apr. 2021, doi: 10.1007/s42979-021-00465-3.

Vancoillie, L., Cockmartin, L., Marshall, N., & Bosmans, H. (2021). The impact on lesion detection via a multi-vendor study: A phantom-based comparison of digital mammography, digital breast tomosynthesis, and synthetic mammography. *Medical physics*, *48*(10), 6270-6292.

Venkatesan, E. v., & Velmurugan, T. (2015). Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology*.

Verma, R., Nagar, V., & Mahapatra, S. (2021). Introduction to supervised learning. *Data Analytics in Bioinformatics: A Machine Learning Perspective*, 1-34.

Wahab, N., Khan, A., & Lee, Y. S. (2019). Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images. *Microscopy*, *68*(3), 216-233.

Wang, J., & Yang, Y. (2018). A context-sensitive deep learning approach for microcalcification detection in mammograms. *Pattern recognition*, *78*, 12-22.

Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., & Jin, Y. (2020). An improved random forest-based rule extraction method for breast cancer diagnosis. *Applied Soft Computing*, *86*, 105941.

Warnecke, A., Arp, D., Wressnegger, C., & Rieck, K. (2020). Evaluating explanation methods for deep learning in security. 2020 IEEE european symposium on security and privacy (EuroS&P).

WHO. (2022a). *Breast fact sheet*. World Health Organization. Retrieved 15 November 2022 from https://bit.ly/3lqFZ6C.

WHO. (2022b). *Gaza strip and West Bank fact sheets*. World Health Organization. Retrieved 15 November 2022 from https:// bit.ly/3uYqW7q.

X. Tang, L. Cai, Y. Meng, C. Gu, J. Yang, and J. Yang, "A novel hybrid feature selection and ensemble learning framework for unbalanced cancer data diagnosis with transcriptome and functional proteomic," IEEE Access, vol. 9, pp. 51659–51668, 2021, doi: 10.1109/ACCESS.2021.3070428.

X. Zhang et al., "Deep learning based analysis of breast cancer using advanced ensemble classifier and linear discriminant analysis," IEEE Access, vol. 8, pp. 120208–120217, 2020, doi: 10.1109/ACCESS.2020.3005228.

Xie, W., Li, Y., & Ma, Y. (2016). Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing*, *173*, 930-941.

Yasaka, K., & Abe, O. (2018). Deep learning and artificial intelligence in radiology: Current applications and future directions. *PLoS medicine*, *15*(11), e1002707.

Yeh, E. D., Jacene, H. A., Bellon, J. R., Nakhlis, F., Birdwell, R. L., Georgian-Smith, D., Giess, C. S., Hirshfield-Bartek, J., Overmoyer, B., & Van den Abbeele, A. D. (2013). What radiologists need to know about diagnosis and treatment of inflammatory breast cancer: a multidisciplinary approach. *Radiographics*, *33*(7), 2003-2017.

Yim, H., Kang, D. K., Jung, Y. S., Jeon, G. S., & Kim, T. H. (2016). Analysis of kinetic curve and model-based perfusion parameters on dynamic contrast enhanced MRI in breast cancer

patients: Correlations with dominant stroma type. *Magnetic Resonance Imaging*, *34*(1), 60-65.
Zeiser, F. A., da Costa, C. A., Zonta, T., Marques, N. M., Roehe, A. V., Moreno, M., & da Rosa Righi, R. (2020). Segmentation of masses on mammograms using data augmentation and deep learning. *Journal of digital imaging*, *33*, 858-868.
Zhang, D., Lv, J., & Cheng, Z. (2020). An approach focusing on the convolutional layer characteristics of the VGG network for vehicle tracking. *IEEE Access*, *8*, 112827-112839.
Zhao, H., Zou, L., Geng, X., & Zheng, S. (2015). Limitations of mammography in the diagnosis of breast diseases compared with ultrasonography: a single-center retrospective analysis of 274 cases. *European journal of medical research*, *20*, 1-7.
Zhou, S. K., Le, H. N., Luu, K., Nguyen, H. V., & Ayache, N. (2021). Deep reinforcement learning in medical imaging: A literature review. *Medical image analysis*, *73*, 102193.
Zhu, Y., & Newsam, S. (2017). Densenet for dense flow. 2017 IEEE international conference on image processing (ICIP).
Zhu, Z., Wang, S.-H., & Zhang, Y.-D. (2023). A survey of convolutional neural network in breast cancer. *Computer modeling in engineering & sciences: CMES*, *136*(3), 2127.

Appendix:

1. Performance of Hybrid Artificial Intelligence (AI) Models without Enhancement techniques:

Table 11: Performance Evaluation of the Original Data (Before image enhancement) Using Different Machine Learning Classifiers, with Respect to Two Deep Learning Models (VGG16, DenseNet121).

Confusion Matrix	Feature Extractor	Classifier	Accuracy	Precision	Recall	F1-Score	AUC	Time (Sec)
Benign Prediction								
Model	**VGG 16**	**Random Forest**	0.846	0.610	0.315	0.460	0.806	36.21
		Gradient Boosting	0.865	0.584	0.338	0.432	0.841	261.52
		SVM	0.829	0.642	0.319	0.418	0.834	52.29
		Logistic Regression	0.848	0.511	0.582*	0.524	0.814	0.80*
	DenseNet 121	**Random Forest**	0.853	0.656*	0.342	0.513	0.836	22.46
		Gradient Boosting	0.869*	0.631	0.381	0.508	0.858*	187.15
		SVM	0.847	0.646	0.332	0.517	0.848	14.97
		Logistic Regression	0.856	0.523	0.543	0.540*	0.837	1.14
Malignant Prediction								
	VGG 16	**Random Forest**	0.850	0.652	0.318	0.471	0.827	10.56
		Gradient Boosting	0.856	0.646	0.398	0.480	0.822	78.38
		SVM	0.846	0.552	0.392	0.465	0.836	13.42

Model		**Logistic Regression**	0.852	0.560	0.460	0.531	0.834	0.2
	DenseNet 121	**Random Forest**	0.867	0.710*	0.369	0.553	0.842	15.58
		Gradient Boosting	0.871	0.625	0.412	0.544	0.841	122.78
		SVM	0.855	0.671	0.486	0.583*	0.855*	11.71
		Logistic Regression	0.872*	0.579	0.564*	0.560	0.843	0.1*

*: The highest value for each hybrid AI model in relation to each evaluation tool.

2. Performance of Hybrid Artificial Intelligence (AI) Models with Enhancement techniques:

Table 12: Performance Evaluation of the Enhanced Data Using CLAHE Algorithm, Using Different Machine Learning Classifiers, with Respect to Two Deep Learning Models i.e. (VGG16, DenseNet121).

Confusion Matrix	Feature Extractor	Classifier	Accuracy	Precision	Recall	F1-Score	AUC	Time (Sec)
Benign Prediction								
Model	**VGG 16**	**Random Forest**	0.952	0.985	0.803	0.814	0.976	4.83
		Gradient Boosting	0.973	0.978	0.867	0.924	0.987	58.25
		SVM	0.975	0.981	0.959	0.841	0.986	4.45
		Logistic Regression	0.985	0.986	0.979*	0.987	0.987	0.13*
	DenseNet 121	**Random Forest**	0.969	0.902	0.881	0.836	0.994	15.43
		Gradient Boosting	0.957	0.985	0.872	0.939	0.992	141.05

		SVM	0.986	0.991	0.946	0.987	0.999*	8.08
		Logistic Regression	0.991*	0.996*	0.978	0.989*	0.999*	0.15
Malignant Prediction								
Model	**VGG 16**	**Random Forest**	0.959	0.988	0.825	0.893	0.992	5.33
		Gradient Boosting	0.976	0.977	0.874	0.933	0.991	43.94
		SVM	0.980	0.989	0.961	0.848	0.994	3.74
		Logistic Regression	0.983	0.960	0.950	0.955	0.996	0.08*
	DenseNet 121	**Random Forest**	0.973	0.988	0.895	0.950	0.996	25.49
		Gradient Boosting	0.982	0.979	0.892	0.947	0.996	219.23
		SVM	0.992	0.992	0.966	0.982	0.998	19.03
		Logistic Regression	0.995*	0.995*	0.995*	0.995*	0.999*	0.22

*: The highest value for each hybrid AI model in relation to each evaluation tool.

3. Validation of Hybrid Artificial Intelligence (AI) Models with Enhancement techniques:

Table 13: Performance Evaluation of the Proposed Hybrid Model on a Real-World Clinical Dataset with Enhancement (400 benign and 400 Malignant) cases, Referring to Confirmed Histopathology.

Confusion Matrix	Feature Extractor	Classifier	Prediction (%)	Accuracy	Precision	Recall	F1-Score	AUC	Time (Sec)
Benign									
	VGG 16	**Random Forest**	90.4	0.833	0.958	0.638	0.766	0.930	1.93

Model		**Gradient Boosting**	91.7	0.925	0.965*	0.777	0.862	0.947	23.30
		SVM	91.7	0.916	0.722	0.893	0.838	0.988	1.78
		Logistic Regression	94.0	0.942	0.872	0.944	0.907	0.986	0.05*
	DenseNet 121	**Random Forest**	95.2*	0.950	0.933	0.824	0.903	0.984	6.17
		Gradient Boosting	92.9	0.933	0.964	0.794	0.871	0.973	56.42
		SVM	94.0	0.933	0.814	0.765	0.866	0.980	3.23
		Logistic Regression	95.2*	0.967*	0.886	0.975*	0.939*	0.999*	0.06
Malignant									
Model	**VGG 16**	**Random Forest**	93.1	0.923	0.816	0.800	0.811	0.948	4.10
		Gradient Boosting	91.7	0.921	0.968*	0.750	0.845	0.974	33.80
		SVM	91.7	0.901	0.706	0.852	0.810	0.991	2.88
		Logistic Regression	94.4	0.937	0.902	0.925	0.914	0.994	0.06*
	DenseNet 121	**Random Forest**	93.1	0.919	0.894	0.729	0.836	0.982	19.61
		Gradient Boosting	93.1	0.928	0.955	0.789	0.862	0.976	168.64
		SVM	94.4	0.928	0.916	0.790	0.857	0.990	14.64
		Logistic Regression	95.8*	0.955*	0.921	0.946*	0.933*	0.998*	0.17

*: The highest value for each hybrid AI model in relation to each evaluation tool.

أسلوب الذكاء الاصطناعي الهجين للكشف المبكر عن سرطان الثدي وتصنيفه من خلال صور أشعة الثدي في فلسطين

إعداد: عمر فايق صادق دراغمه

المشرف: د. رضوان قصراوي

الملخص

يعد سرطان الثدي مصدر قلق صحي عالمي كبير، والتشخيص المبكر والدقيق أمر بالغ الأهمية لتحسين نتائج المرضى ومعدلات البقاء على قيد الحياة. ومع ذلك، على الرغم من التقدم في التكنولوجيا الطبية وتقنيات الفحص، لا يزال خطأ التشخيص يمثل تحديًا مستمرًا في الكشف عن سرطان الثدي. تبحث هذه الدراسة في استخدام نماذج الذكاء الاصطناعي الهجين (AI) التي تجمع بين تقنيات التعلم العميق والتعلم الآلي للتنبؤ بسرطان الثدي الحميد والخبيث من صور الثدي بالأشعة السينية. تبدأ الدراسة باستخدام نماذج الشبكة العصبية التلافيفية المدربة مسبقًا، وهي VGG16 وDenseNet121، لاستخراج الميزات من صور تصوير الثدي بالأشعة السينية. تم تدريب نماذج التعلم العميق هذه على مجموعات بيانات كبيرة وتعلمت كيفية تحديد الأنماط والميزات المختلفة داخل الصور. ومن خلال استخراج هذه الميزات من تصوير الثدي بالأشعة السينية، يمكن للنماذج التقاط معلومات مهمة ذات صلة بتصنيف سرطان الثدي. يتم بعد ذلك استخدام الميزات المستخرجة لتدريب العديد من مصنفات التعلم الآلي، بما في ذلك الانحدار اللوجستي، وآلات ناقلات الدعم، والغابات العشوائية، ونماذج تعزيز التدرج. تتعلم هذه المصنفات التعرف على الأنماط وإجراء التنبؤات بناءً على الميزات المستخرجة.

لتقييم أداء نماذج الذكاء الاصطناعي الهجين، تم إجراء الدراسة على ثلاث مراحل. في المرحلة الأولى، يتم استخدام صور أشعة الثدي الأصلية للتصنيف. في المرحلة الثانية، يتم تحسين صور أشعة الثدي باستخدام تقنيات المعالجة المسبقة المختلفة للصور وتحسينها. وأخيراً، في المرحلة الثالثة، يتم اختبار النماذج على صور اشعة الثدي جديدة لتقييم قدراتها على التعميم. لتحسين صور أشعة الثدي، يتم تطبيق العديد من تقنيات معالجة الصور. وتشمل هذه المعالجة المسبقة للتآكل المورفولوجي، ومعادلة الرسم البياني التكيفي المحدود التباين (CLAHE)، وتحسين حافة Laplacian of Gaussian (LoG)، والإخفاء غير الواضح. تهدف هذه التقنيات إلى تحسين رؤية الهياكل والميزات المهمة داخل الصور، مما يسهل على نماذج الذكاء الاصطناعي إجراء تنبؤات دقيقة.

وفي المرحلة الثانية، عند التنبؤ بالحالات الحميدة من خلال صور أشعة الثدي المحسنة، يحقق مصنف الانحدار اللوجستي المزود بميزات DenseNet121 أداءً رائعًا. إنها تحقق أعلى دقة تبلغ 0.991، ودقة تبلغ 0.996، ودرجة F1 تبلغ 0.989، وAUC تبلغ 0.999. تعمل أيضًا آلة ناقل الدعم المزودة بميزات DenseNet121 بشكل جيد، بدقة تبلغ 0.986 وAUC تبلغ 0.999. يوضح نموذج الانحدار اللوجستي مع ميزات VGG16 أسرع وقت تنبؤي، ويتطلب 0.13 ثانية فقط. وبالمثل، في التنبؤ بالحالات الخبيثة من الصور المحسنة، يتفوق مصنف الانحدار اللوجستي مع ميزات DenseNet121 بأعلى دقة تبلغ 0.995، ودقة تبلغ

0.995، واستدعاء 0.995، ودرجة F1 تبلغ 0.995، وAUC تبلغ 0.999. تتبع آلة ناقل الدعم المزودة بميزات DenseNet121 بدقة تبلغ 0.992 وAUC تبلغ 0.998. يحافظ نموذج الانحدار اللوجستي مع ميزات VGG16 على وقت تنبؤي سريع، حيث يستغرق 0.08 ثانية فقط. وتوضح الدراسة أن صور التصوير الشعاعي للثدي المحسنة في المرحلة الثانية تتفوق باستمرار على صور الاختبار الأصلية والجديدة في المرحلتين الأولى والثالثة، على التوالي. وهذا يؤكد على التأثير الكبير لتقنيات المعالجة المسبقة للصور وتحسينها على القدرات التنبؤية لنماذج الذكاء الاصطناعي الهجين. تسلط النتائج الضوء على إمكانية الجمع بين التعلم العميق لاستخراج الميزات والتعلم الآلي للتصنيف في تحقيق دقة عالية ودقة واسترجاع ودرجات F1 وقيم المساحة تحت المنحنى للتنبؤ بسرطان الثدي الخبيث من صور تصوير الثدي بالأشعة السينية.

في الختام، توضح الدراسة إمكانات نماذج الذكاء الاصطناعي الهجين التي تجمع بين تقنيات التعلم العميق والتعلم الآلي للتنبؤ بسرطان الثدي الحميد والخبيث من صور الثدي بالأشعة السينية. يؤدي دمج التعلم العميق لاستخراج الميزات والتعلم الآلي للتصنيف، إلى جانب المعالجة المسبقة للصور وتحسينها، إلى تحسين الدقة والأداء. هذه التطورات لديها القدرة على تعزيز الكشف عن سرطان الثدي، مما يؤدي في النهاية إلى نتائج أفضل للمرضى ومعدلات البقاء على قيد الحياة.